Transcriptional networks in at-risk individuals identify signatures of type 1 diabetes progression

Louis-Pascal Xhonneux[1,2], Oliver Knight[1, 2], Åke Lernmark[4], Ezio Bonifacio[5], William A. Hagopian[6], Marian J. Rewers[7], Jin-Xiong She[8], Jorma Toppari[9,10], , Hemang Parikh[13], Kenneth G.C. Smith[1,2], Anette-G. Ziegler[11], Beena Akolkar[12], Jeffrey P. Krischer[13] and Eoin F. McKinney*[1,2,3]

1 Cambridge Institute of Therapeutic Immunology and Infectious Disease, Jeffrey Cheah Biomedical Centre, Cambridge, CB2 0AW, U.K.
2. Department of Medicine, University of Cambridge School of Clinical Medicine, Cambridge, CB2 0QQ, U.K.
3. Cambridge Centre for Artificial Intelligence in Medicine
4. Department of Clinical Sciences, Lund University/CRC, Skåne University Hospital, Malmo, Jan Waldenströms gata 35, Malmö, Sweden
5. Center for Regenerative Therapies, Technische Universität Dresden, Fetscherstraße 105
01307, Dresden
6. Pacific Northwest Research Institute, 720 Broadway, Seattle, WA 98122, U.S.A.
7. Barbara Davis Center for Childhood Diabetes, University of Colorado, 1775 Aurora Ct, Aurora, CO 80045, U.S.A.
8. Center for Biotechnology and Genomic Medicine, Medical College of Georgia, Augusta University, 1462 Laney Walker Blvd. Augusta, GA 30912, U.S.A.
9. Department of Pediatrics, Turku University Hospital, Kiinamyllynkatu 4-8, 20521 Turku, Finland
10. Institute of Biomedicine, Research Centre for Integrative Physiology and Pharmacology, University of Turku FI-20014 Turun Lyliopisto, Finland
11. Institute of Diabetes Research, Helmholtz Zentrum München, and Klinikum rechts der Isar, Technische Universität München, and Forschergruppe Diabetes e.V., Arcisstraße 21, 80333 München, Germany
12. National Institute of Diabetes & Digestive & Kidney Diseases, 9000 Rockville Pike
Bethesda, MD 20892, U.S.A.
13. Health Informatics Institute, Morsani College of Medicine, University of South Florida, Tampa FL 33612, U.S.A.
*corresponding author

Overline: METABOLISM

One-Sentence Summary

Transcriptomic analysis identifies progression signatures facilitating individual prediction of T1D risk from 18 months of age in independent cohorts.

Abstract

Type 1 diabetes (T1D) is a disease of insulin deficiency that results from autoimmune destruction of pancreatic islet beta cells. The exact cause of T1D remains unknown, although asymptomatic islet autoimmunity lasting from weeks to years prior to diagnosis raises the possibility of intervention before the onset of clinical disease. The number, type, and titer of islet autoantibodies are associated with long-term disease risk but do not cause disease and robust early predictors of individual progression to T1D onset remain elusive. The Environmental Determinants of Diabetes in the Young (TEDDY) consortium is a prospective cohort study aiming to determine genetic and environmental interactions causing T1D. Here, we analysed longitudinal blood transcriptomes of 2013 samples from 400 individuals in the TEDDY study prior to both T1D and islet autoimmunity. We identified and interpreted age-associated gene expression changes in healthy infancy and age-independent changes tracking with progression to both T1D and islet autoimmunity, beginning before other evidence of islet autoimmunity is present. We combined multivariate longitudinal data in a Bayesian joint model to robustly predict individual risk of T1D onset, validating the association of a natural killer cell signature with progression as well as the model's predictive performance on an additional 356 samples from 56 subjects in the independent Type 1 Diabetes Prediction and Prevention (DIPP) study. Together our results indicate that T1D is characterised by early and longitudinal changes in gene expression, informing the immunopathology of disease progression and facilitating prediction of its course.

Introduction

An autoimmune pathogenesis for T1D is indicated by strong genetic association of HLA and other immune variants (*1*) along with progressive development of pancreatic islet beta cell autoantibodies (IAb) (*2*), although metabolic, microbial, and dietary factors also contribute (*3*). Genetic risk scores can identify those at highest risk of developing disease (*4*) in whom peak onset occurs in early childhood, around 1-2 years of age (*5*). Studying early events associated with T1D progression is challenging given the need to identify at-risk subjects and to sample before evidence of autoreactivity begins. Previous studies have typically been cross-sectional investigations of small numbers of subjects after the onset of islet autoimmunity or T1D. Expansions of islet antigen-reactive T cells have been documented in the blood of both T1D cases and healthy controls (*6*) whereas age-dependent changes in immune phenotype at diagnosis (*7*) have underlined the complexity of studying disease in the context of a developing immune system. Within the pancreas, longitudinal single-cell RNAseq analysis has indicated complex, dynamic changes in infiltrating immune cell populations (*8*). Although comparable study of human tissue is not possible, pseudotime mass cytometry of pancreatic tissue from T1D-diagnosed donors has illustrated an influx of cytotoxic and helper T cells associated with reductions in beta cell mass (*9*).

Longitudinal metabolomic (*10*) and microbiomic (*11, 12*) analyses prior to onset of autoimmunity or T1D diagnosis have indicated marked age-dependent effects but not a clean association with autoimmunity or disease progression. A viral trigger for pancreatic autoimmunity has long been proposed (*13*) and has been supported by both early transcriptional signatures of type 1 interferon response (*14*) and prolonged enteroviral shedding (*15*) prior to islet autoimmunity. Enteroviruses can directly infect pancreatic beta cells and have been linked to a natural killer cell insulitis (*16*). Early changes in blood gene expression (*17, 18*) prior to islet autoimmunity are likely to be driven by changes in relative proportions of constituent blood cells(*19*), although their interpretation is confounded by a lack of data describing how such changes develop in healthy infancy.

T1D onset occurs in the dynamic context of a maturing immune system: changes observed in children developing autoimmunity must be carefully related to healthy developmental changes to focus our attention on factors that initiate and propagate autoreactive responses. Recently, systems immunology studies of immune development during infancy have indicated early, stereotyped changes in blood cell and protein composition (*20*) and dynamic changes in the gut microbiome related to breastfeeding patterns, islet autoimmunity (*11, 12*), and onset of childhood inflammatory disease (*21*). Such high-throughput analyses create the potential for identification of previously unsuspected pathways associated with disease initiation and progression to improve understanding of disease biology and aid the building of predictive models to suggest targeted therapies (*22*).

The TEDDY study aims to identify gene-environment interactions causing T1D in high-risk infants participating in serial prospective sample collection and monitoring every three to six months from birth to age 15 years (*23*). By combining genetic risk stratification and prospective follow-up it has been possible to collect samples and data from infants prior to development of both islet autoimmunity and T1D onset, facilitating systematic analysis of early and longitudinal changes associated with the later development of disease. This can facilitate both an improved understanding of the pathogenetic mechanism of human islet autoimmunity and early prediction of its subsequent course. Although the presence of IAb indicates the development of islet

autoimmunity and hence long-term risk of T1D development (*2*), validated predictive markers have not been able to report serial, individual risk over a near-term horizon such as is necessary to inform clinical decision making (*24*).
Here, we undertook transcriptional network analysis of gene expression microarray data using a nested case:control cohort (*25*) from the TEDDY study to identify early longitudinal changes in whole blood gene expression in healthy infancy or tracking with progression to both islet autoimmunity and T1D. We also built a predictive model incorporating multivariate longitudinal features including gene expression and islet autoantibodies to estimate individual risk of T1D progression.

Results

A dynamic landscape of whole-blood gene expression during infancy

We analysed data from nested, matched case control cohorts (*25*) comprising 2013 whole blood transcriptomes sampled longitudinally from 401 individuals, divided into those developing islet autoimmunity or T1D (Fig. 1A, B, fig. S1-2). We applied transcriptional network analysis (*26*) to identify groups of coexpressed genes (modules) based solely on patterns of transcription in the data (Fig. 1C, D, fig. S1B). We first constructed independent transcriptional networks for both cases and controls, observing closely matched coexpression patterns in each with no evidence of disease-specific modules (fig. S1B): genes coexpressed in cases were found to be similarly coexpressed in the control cohort and vice versa (Fig. 1C, D). As genes in a module are by definition coexpressed, they can be summarised by a single profile known as an eigengene (*27*). As coexpression patterns were preserved in disease and control groups, we constructed a coexpression network based on all samples considered together (fig. S1B) and applied linear mixed modelling to whole blood modular eigengenes from this combined network (table S1). We found that a substantial proportion of modules (23/85, 27%) demonstrated significant temporal changes during infancy (FDR < 5%, Fig. 1E). Patterns of gene coexpression are highly conserved (*28*) and modular signatures can be interpreted by screening their composite genes for enrichment of well-defined gene expression signatures in external relational databanks and public repositories (fig. S1) (*29, 30*). In this way, the biological meaning of the identified patterns of gene coexpression can be interpreted, with the caveat that repositories typically include samples and data from adults rather than infants. We compared each age-associated module (Fig. 1E) to the largest compendium of cell- and tissue-specific transcriptomic signatures (ARCHS4 repository (*30*)), identifying frequent modular enrichment for cell type-specific transcripts (17/23, 74%), upstream kinases (22/23, 96%) and transcription factors (22/23, 96%) (fig. S3) including progressive reductions in stem cell-specific transcripts (Fig. 1F) along with B cell- and neutrophil-specific transcripts and associated transcription factors (fig. S3).
We also undertook ‘digital cytometry’, using a deconvolution approach comparing whole blood transcriptional profiles to immune cell-specific transcripts to estimate cell type frequencies (Fig. 1G) (*31*). This confirmed the association of the majority of modular signatures identified with the frequency of circulating immune cell phenotypes (Fig. 1G). Individual modular signatures correlated with multiple deconvoluted cell frequencies as expected, as cell percentages often vary together during an immune response.
Multiple modular signatures mapped to the same cell type despite showing distinct longitudinal trajectories, suggesting they cannot be explained by changes in the

proportion of circulating cell types in blood alone and likely reflect more subtle phenotypic differences than can currently be mapped by deconvolution or enrichment strategies. Together these data show marked gene expression changes during healthy infancy, highlighting the dynamic context in which autoimmune diseases such as T1D occur.

Distinct disease-specific transcriptional signatures in T1D subgroups

We next asked whether longitudinal changes in modular patterns of gene expression tracked with progression towards T1D onset. As children get older, they inevitably approach the time of disease onset. Consequently, it is necessary to ensure that the longitudinal changes seen in cases are not simply age-associated changes expected to occur in healthy children. To overcome this, we identified modular eigengenes correlating with time to T1D in cases and, as the study design included age-matched healthy samples, compared these to the longitudinal changes seen in infants who did not progress to disease (Fig. 1H, I). Although 35 modular signatures significantly correlated with T1D progression (FDR<5%), none were specific to T1D cases and all showed stronger association with healthy ageing in matched controls (Fig. 1I), highlighting the importance of placing longitudinal changes into the dynamic immune context of early infancy.

T1D is thought to be a heterogeneous condition, however evidence supporting the existence of disease subgroups has proven elusive (*32, 33*). Those developing autoantibodies to insulin (IAA) show earlier and more rapid progression to both additional IAb and to T1D than those initially developing IAb to the other major pancreatic autoantigen, GAD (glutamic acid decarboxylase) (*2*). However, it remains unclear whether this observation reflects a distinct immunopathology, or simply autoimmunity occurring at younger age. We next stratified the T1D cohort by target of first appearing autoantibody, comparing modular gene expression changes in IAA-first and GADA-first subgroups with their matched controls as before. Whereas disease-specific longitudinal changes were not apparent in T1D on the whole (Fig. 1I), distinct gene expression signatures showed clear age-independent association with time to T1D onset in IAb subgroups (Fig. 2A, B). Amongst IAA cases, one dominant signature (IAAsig) showed an early increase in expression with a later secondary increase prior to diagnosis (Fig. 2C), a pattern not seen in most matched controls. The nested study design allowed for IAb seroconversion to occur in the T1D control group (*25*), provided T1D progression did not occur. Closer analysis of the control group demonstrated that IAAsig expression showed a similar early rise in seroconverting controls, falling off with advancing age where they were maintained in those progressing to T1D (Fig. 2C inset, $IAb^{pos}T1D^{neg}$ vs. $IAb^{neg}T1D^{neg}$). By contrast, seronegative controls showed no comparable rise in early IAAsig expression (Fig. 2C inset). In GADA cases, a group of four closely correlated signatures termed together as GADAsig, (Fig. 2B) showed an age-independent decrease towards T1D onset, a pattern absent from matched controls and occurring closer to diagnosis in contrast to the earlier increases in the IAAsig (Fig. 2C, D). We compared mixed models to ensure that observed gene expression changes were independent of additional clinical covariates including sex, ethnicity, and HLA risk group (table S2) alongside mode of birth delivery and patterns of maternal breastfeeding (fig. S4C-F). This analysis indicated an independent and significant association of both gene expression signatures and sex, but not the other variables, to the models (table S3). We therefore took these covariates forward into predictive model building as described below.

Biological interpretation of T1D-specific signatures

We next performed enrichment analysis of both IAAsig and GADAsig (table S4) against relational databases (ARCHS4 (*30*) and DICE (*34*)) providing the largest and most granular coverage of immune cell-specific gene expression, and also correlated each module's eigengene against cell subset frequencies estimated through deconvolution analysis (*31*). IAAsig genes showed strong, specific enrichment for natural killer (NK) cell-specific transcripts (Fig. 2E, F), transcription factors, and kinases (fig. S5) and correlated with deconvoluted percentage of NK cells, and to a lesser but still significant extent with CD4+ memory T cells (Fig. 2G). Genes comprising GADA-specific modules were enriched for transcripts shared by both blood and brain tissue but not clearly with a single cell type (Fig. 2H, I, fig. S5). However, comparison with deconvolved cell type frequencies indicated strongest associations with reduced percentage of CD4+ memory T cells and NK cells, with a relative increase in an activated NK phenotype (Fig. 2J). This observation suggests the early stages of T1D pathogenesis are associated with different immune cell trajectories that involve similar immune cell types, namely NK and CD4+ memory T cells, depending on the pattern of initial IAb seroconversion.

To investigate potential means of therapeutically modulating IAAsigs and GADAsigs, we compared each to an integrated repository of drug response data (the Harmonizome (*35, 36*)), that links functional associations between genes and proteins based on collated genomic data including physical associations, knockout or knockdown phenotypes, and response to drug treatment. We screened IAAsig genes against all 352 'druggable' targets (*30*) linked to 20883 genes and identified a single candidate G-protein coupled receptor (GPR171, Fig. 2K) as a potential controller of IAAsig genes. That is, GPR171 was not itself part of the modular signature but instead predicted to be functionally associated with genes comprising it. We confirmed NK cell -expression of the IAAsig and GPR171 at both mRNA (Fig. 2L) and protein abundance (Fig. 2M) and demonstrated that a specific inhibitor of its signalling (Inh) could attenuate both GPR171 expression and NK cytotoxicity in an *in vitro* killing assay (Fig. 2N). Together, these data demonstrate that distinct cell-specific gene expression changes characterise progression to disease onset in subgroups of patients with T1D defined by their sequence of IAb seroconversion.

Transcriptional signatures associated with islet autoimmunity

Next, we asked whether specific changes in gene expression occur around the onset of islet autoimmunity (IAb seroconversion), rather than tracking with progression to disease onset (Fig. 3A). Amongst 50 modular signatures showing significant association with islet autoimmunity onset (Fig. 3B, C), a dominant signature associated with seroconversion in both subgroups. This signature (IAsig) was common to both GADA and IAA subgroups and, although it showed significant dynamic changes in matched healthy controls, there were more marked and sustained reductions in infants progressing to IAb seroconversion (Fig. 3D, E). Interrogation of this islet autoimmunity signature (table S5) revealed strong enrichment for B cell-specific transcripts, kinases, and transcription factors (Fig. 3F). Six additional signatures (Fig. 3B, C) showed a weaker but still significant age-independent association with islet autoimmunity onset, again in both GADA and IAA subgroups. The second-most strongly associated signature was the same NK-enriched signature that associated with T1D onset in IAA, which here increased towards islet

autoimmunity onset but showed no change over time in the control group (Fig. 3G, H). Together these data indicate that longitudinal changes in NK- and B cell-associated gene expression track with progression towards the onset of islet autoimmunity with NK associated changes also tracking with progression to disease in the IAA subgroup.

Validation of T1D-specific longitudinal gene expression changes

We next sought to validate our findings in an independent cohort of IAb and T1D cases. The DIPP cohort (*37*) is a prospective, population study of incident islet autoimmunity and T1D with a comparable nested case:control design although with sampling commencing in slightly older children (*18*). We undertook an independent network transcriptomic analysis of 356 DIPP samples from 58 individuals, again comparing dynamic changes in modular gene expression to both islet autoimmunity (Fig. 3I) and T1D onset (Fig. 3J). In this smaller cohort, an NK cell-enriched signature (directly comparable to that identified in TEDDY, fig. S6, table S6) was the only gene expression pattern that showed a significant association with both IAb seroconversion and T1D progression which did not similarly change in matched controls (Fig. 3I, J). A B cell signature comparable to the IAsig seen in TEDDY (Fig. 3G, H) was significantly associated with both clinical endpoints, but in the DIPP cohort showed a comparable association with sampling age. Similar to what was observed in the TEDDY cohort, the NK signature increased towards T1D onset with a later decline in matched controls who did not go on to develop disease (Fig. 3I, J). Together, these data confirm an independent association of an NK cell-enriched transcriptional signature with both IAb seroconversion and rate of progression to T1D, validating the finding in the larger TEDDY discovery cohort.

Gene expression in early infancy associates with rate of disease progression

We next sought to investigate whether whole blood gene expression changes in early infancy, prior to demonstrable evidence of islet autoimmunity, were related to later risk of progression towards T1D. For this 'snapshot' of early risk, we identified the earliest samples available within the case:control cohorts, comprising 288 samples from 288 individuals, all taken prior to seroconversion and with >85% taken within the first 12 months of life (Fig. 4A, B). In this cross-sectional analysis, we constructed a gene coexpression network and looked for evidence of association with both disease risk (outcome $T1D^+$ vs $T1D^-$) and rate of subsequent progression towards T1D. As longitudinal changes were not being considered, we used regression analysis to adjust for variable sampling age and sex. At this early timepoint four modules were associated with the rate of subsequent progression towards T1D (Fig. 4C, D). Both enrichment and deconvolution correlation analyses of these modules indicated specific excess of B lymphoblast and monocyte expressed transcripts respectively with the latter also enriched for TNF and complement pathway signalling (Fig. 4E, table S7, figS6H, I). These data indicate that early high expression of a TNF-enriched monocyte signature and early low expression of a B lymphoblast signature were associated with slower progression to T1D onset (Fig. 4C, D). No modular signatures were significantly different between T1D and healthy control groups, although the TNF-enriched monocyte signature that associated with protection against T1D was markedly higher in infants who later seroconverted without developing T1D (IA^+T1D^-, Fig. 4F). We also observed that the same pattern of monocyte/TNF associated transcripts showed significantly lower expression in children within a window of 12

months prior to diagnosis of T1D (Fig. 4G), a finding that was also validated in the DIPP cohort (Fig. 4H).

Previous analyses have identified a type 1 interferon response signature expressed in at-risk children prior to antibody seroconversion and associated with previous respiratory infections (*14*). Such a signature was clearly visible in the TEDDY cohort (Fig. 4I) although it was not associated with risk of T1D (Fig. 4C) or rate of progression to T1D (Fig. 4J). However, the IFN response signature conformed to a pattern of transient 'spikes' of expression, likely following infectious triggers, most frequently observed in the 12 months prior to disease onset (Fig. 4J, K) and to an extent that exceeded those seen in age-matched control samples (Fig. 4L).

Prediction of individual T1D risk using longitudinal data

Recently, statistical learning methods have improved our ability to integrate baseline covariates, longitudinal data. and clinical endpoints to estimate instantaneous event hazards (*38*). Current T1D prediction methods stratify disease risk by number of IAb present, indicating cohort level risk over a horizon of many years, making it difficult to incorporate this information into treatment pathways or to enable recruitment into clinical trials of targeted therapy (Fig. 1B). Individual risk prediction over a short time horizon is necessary to guide clinical decisions and preventive therapy trials (*24*). We therefore aimed to build a predictive model that could estimate, with an indication of uncertainty, the near-term hazard of T1D for an individual, dynamically updating that prediction as additional data becomes available. We used a multivariate Bayesian joint model (*39*) to combine baseline stratification (using a Cox proportional hazards model) with longitudinal variables such as IAb type, status, timing, and gene expression, returning an event hazard with associated confidence bounds. As our earlier analysis of comparative mixed models had demonstrated independent association of both gene expression signatures and sex (table S3) with gene expression signatures depending on the sequence of serial IAb seroconversion, we incorporated these covariates into a joint model (Fig. 5A). We chose to build and test the performance of three distinct models in each of two clinical scenarios. Predictive performance (ROC AUC and prediction error) was estimated using 10-fold cross-validation on the discovery cohort (TEDDY, Fig. 5B) and on an independent validation dataset (DIPP). The three models each included baseline stratification by sex and either (i) IAb status over time ($IAb^{+/-}$), (ii) maximal IAb information (IAb specificity [IAA, GADA, IA-2A], timing of IAb seroconversion, sequence of IAb seroconversion specificity and associated interaction effects) or (iii) maximal IAb information along with longitudinal gene expression data (the eigenvalues of the IAA or GADA signatures). The two clinical scenarios tested were serial prediction over a fixed future horizon of 12 months using all cumulative data available at each timepoint (mimicking a child being followed up over time, Fig. 5C-F), and prediction at 1.5 years of age over a serially increasing future horizon (mimicking prediction with early limited data, Fig. 5C-F).

Using islet autoimmunity status alone – the scenario most comparable to current methods (*24*) - had modest predictive accuracy in both clinical scenarios (model i, Fig. 5C). The inclusion of maximal IAb information (model ii) allowed for robust prediction in the first clinical scenario (serial prediction over a fixed horizon of 12 months: Fig. 5D, left panel: ROC AUC>0.9, PE <10%) with only modest improvement by inclusion of longitudinal gene expression signatures (model iii, Fig. 5E). Although the performance of serial IAb in this scenario is an improvement in

T1D prediction, it is also apparent that making predictions close to diagnosis (in this scenario predicting 12 months ahead) is supported by using IAb data. By contrast, and consistent with the importance of early gene expression measures, gene expression signatures supported model performance (model iii) more strongly in the second clinical scenario, where predictions were made early over a serially extending time horizon. This was particularly apparent with prediction over the first few years of life, when the majority of infant T1D cases occurred (Fig. 5D and E). These data show that, although the presence of islet autoantibodies is associated with disease risk (*40*), incorporating information on serial changes in the type, number and timing of seroconversion can facilitate T1D risk prediction at an individual level over a time horizon short enough to facilitate changes in clinical monitoring or therapeutic trials. Gene expression measures provided greatest support for prediction when measured early (up to 18 months) to predict T1D risk over a longer time horizon (up to 5 years in this dataset).

Discussion

Together our data describe dynamic changes in the infant blood transcriptome and show that patterns of islet antibody seroconversion define subgroups of T1D with both distinct rates of progression and distinct age-independent gene expression signatures associated with time to disease onset. Amongst healthy infants we observed extensive longitudinal changes in gene expression over the first five years of life, highlighting the dynamic immune context in which early islet autoimmunity develops. This observation reinforces the importance of taking such changes into account when seeking to differentiate disease-specific changes from those reflecting 'healthy' immune development.

On taking age-associated changes into account, we observed specific, longitudinal changes in gene expression tracking with progression towards both islet autoimmunity and T1D onset. Distinct changes were associated with T1D progression in subgroups defined by the target of initial seroconversion. The two dominant serospecificities at onset of islet autoimmunity (IAA and GADA) have been proposed as distinct disease 'endotypes', with the former developing earlier and showing faster progression towards T1D onset (*41*). Consistent with a distinct pathogenetic mechanism underpinning this stratification, we observed that distinct transcriptional signatures tracked with progression to T1D onset in subgroups defined by the specificity of the first appearing islet autoantibody (IAA or GADA). Earlier changes tracking progression to onset of islet autoimmunity were similar in both groups, however. We identified an NK cell-based signature that increased in expression with progression towards both islet autoimmunity and T1D in IAA-first individuals. The same signature was similarly seen to associate with time to islet autoimmunity but not T1D onset in GADA-first subjects. Association of a very similar NK cell signature was validated in an independent analysis of longitudinal samples from the DIPP study. However, further work is required to understand the mechanism underlying this association. NK cells have a complex relationship to autoimmunity and may function as either effector cells contributing to tissue damage, or as regulators of immunopathology(*42*). Differences in NK cell phenotype have been described after T1D diagnosis (*43*) – although accompanied by many other late differences (*44*) - whereas changes in their number and phenotype have been variably linked to either aggressive insulitis (*45*) or protection from it (*46*) in animal models. Perhaps the most likely explanation for the observed association with T1D progression here is that a viral trigger results in altered NK cell phenotype that tracks with progressive insulitis.

Persisting enteroviral infection has been associated with T1D progression (*15*) and is known to infect pancreatic beta cells, inducing early NK infiltration and cytolysis in animal models (*47, 48*). Our data indicate a prominent and specific role for NK cells in the development and progression of autoimmunity and T1D in humans, beginning at the earliest stages and tracking longitudinally with rate of progression rather than simply differentiating those who already have disease from those who do not. Although it is difficult to further refine the source of the NK-specific transcriptional signature using whole blood data, it is unlikely that it simply reflects a relative expansion of peripheral NK cells, as evidenced by our identification of other NK cell-enriched modular signatures that did not associate with disease progression. Perhaps the most likely explanation for the observed NK association with T1D progression is that a viral trigger results in altered NK cell phenotype that tracks with progressive insulitis. Persisting enteroviral infection has been associated with T1D progression (*15*) and enteroviruses are known to infect pancreatic beta cells, inducing early NK infiltration and cytolysis in animal models (*47, 48*). An NK-predominant insulitis has also been observed in pancreas from diabetic organ donors with Coxsackie B4 enteroviral infection (*16*) and is consistent with an autoreactive effector role for NK in T1D pathogenesis. However, NK cells may also function to regulate T cell-mediated immunity during persistent viral infection (*49*). A limitation of the current study is that, while the association of an NK transcriptional signature is validated and may be used for predicting progression, the mechanism linking NK cells to insulitis requires further investigation.

Despite these limitations, we also demonstrate that a disease-relevant transcriptional signature can serve as the starting point for further mechanistic understanding and novel therapeutic approaches. By screening the NK-enriched T1D progression-associated signature against collated genomic information from many sources, we predicted and confirmed *in vitro* that inhibition of a poorly characterised G-protein coupled receptor (GPR171*,* previously known for its role in controlling satiety signalling in the hypothalamus (*50*)) was capable of suppressing NK cytolytic protein expression.

We also undertook a systematic network analysis of age-independent blood transcription, prior to IAb emergence with the large majority of samples taken during the first year of life. In these earliest samples we observed reciprocal association of B lymphoblastic and TNF-enriched monocytic signatures that associated with the subsequent rate of progression to T1D. Although these signatures were not different between those later developing T1D and matched controls they were specifically increased in subjects who progressed later to islet autoimmunity without developing T1D ($IAb^{+}T1D^{-}$). Although it is tempting to speculate that this evidence supports a protective role for early inflammatory signals – such as proposed by proponents of the 'hygiene hypothesis' (*51*) – studies in animal models have highlighted the complexity of altered TNF signalling with evidence for distinct roles at different disease stages (*52, 53*). However, the validated association of increased expression of this signature in T1D-protected individuals despite islet autoimmunity may help inform interventional study design (*54*).

Previous hypothesis-driven analyses of early gene expression changes identified an increase in type 1 interferon (IFN1) signalling in pre-T1D children linked to history of recent infection (*14*). A comparable signature was apparent in the TEDDY cohort and showed transient elevation in 'spikes' consistent with response to an infectious stimulus (and quite different to the chronic, progressive increase seen in the NK signature). Greater IFN1-induced gene expression was observed during the 12 months

preceding T1D onset compared with age matched controls but was not associated with the rate of progression to islet autoimmunity or T1D. This is consistent with a role for IFN1 signalling – and perhaps viral infections that provoke transient IFN1 elevations - in modifying disease progression. However, as with NK cells, evidence from animal models shows that IFN signalling may play either a role in promoting T cell mediated insulitis(*55*) or in protecting beta cells from NK cell mediated attack(*48*).

Last, we sought to incorporate the longitudinal measurement of immune traits – both gene expression and IAb – into a predictive model that could provide an estimate of an individual's T1D risk and the confidence of that estimate. Long-term risk of T1D (over the subsequent 10-15 years) can currently be informed by the extent of IAb seropositivity. However, for a predictive model to impact on clinical decision making – whether by altering the frequency of clinical review to monitor for severe complications such as diabetic ketoacidosis (*56*) or by facilitating early intervention studies (*24*) – it is necessary to obtain a robust estimate of near-term risk of T1D onset. We therefore sought to build a predictive model that could estimate individual T1D risk in two specific scenarios: either making an early prediction (at 18 months) over a longer horizon (5 years), or by using cumulative data to make serial predictions over the subsequent 12 months. To test the ability of both baseline and longitudinal measures to inform this prediction, we built a Bayesian joint model incorporating either Ab status alone, or with more extensive IAb features (serospecificity, timing, and interaction of IAb development) with or without gene expression signatures. We included stratification by sex (as this was the only other covariate demonstrating independent association with progression rate), but intentionally excluded HLA stratification (despite a demonstrated association with progression (*57*)) to facilitate extrapolation between global populations with distinct HLA distributions. This approach allowed direct comparison between both simple and more complex models, aiming to establish optimal prediction with the simplest approach requiring as few measurements as possible. With predictions made over a short horizon of 12 months, the model with extensive IAb features outperformed standard prediction using IAb status alone and gained little support from including gene expression data: this is consistent with observations that IAb are often positive within 12 months of diagnosis(*58*) and our model supported robust prediction of T1D progression in this scenario. However, it is an onerous task to repeatedly sample children at such an early age to obtain longitudinal data on timing and sequence of seroconversion specificities. We therefore tested a second scenario, using data only from the first 18 months and making predictions progressively further ahead. Predicting from this earlier timepoint – arguably a more feasible clinical scenario given the reduced sampling requirement – showed a benefit of gene expression signatures in addition to IAb measures with robust performance on both cross-validation and independent validation cohort testing.

The current study identified extensive, longitudinal changes in the whole blood transcriptome occurring during both healthy infancy and progression to T1D. This finding has been made possible through assiduous prospective collection of samples by the TEDDY consortium. We show here that these changes can be both interpreted and used to inform prediction of T1D risk from an early age. Extensive sampling at an early age is facilitated by the simplicity of whole blood collection. However, this method also limits the biological interpretation of modular signatures identified. The modular signatures identified here are dominated by cell-subset specific transcripts, with both module enrichment and deconvolution methods in broad agreement. Each method can identify the likely cellular source of a transcriptional signature but, when

that signal is derived from a mixed cell population like peripheral blood, it is more difficult to pin down the cell-intrinsic pathways responsible for that change in gene expression. Improved methods for deconvolution may help to address this problem (*59*) but require robust validation against concurrently sampled cell intrinsic transcriptomes. Transcriptional profiling of sorted cell populations (*60*) or single cell profiling (*61*) methods can similarly overcome this limitation, but they inevitably result in sampling of a much smaller cohorts. It is clear from our analyses that enrichment and deconvolution approaches can be complementary. As deconvoluted cell subset proportions may vary together, for example increasing together during an inflammatory response, it is expected that a transcriptional signature may correlate with multiple cell subset proportions, making it harder to define the source of that signal through deconvolution alone. Enrichment is not similarly encumbered by this problem, relying instead on co-expressed features within the module itself for interpretation although it is inevitably constrained by the availability of external signatures for enrichment analysis.
We have demonstrated and validated an association of NK cell gene expression signature with T1D progression. It remains to be determined whether this change reflects a causal contribution to T1D related immunopathology or a host response to an infectious trigger, or both. An answer to this fundamental question will require further analyses and more detailed investigation of prospective data and samples.

Longitudinal measurement of gene expression patterns in infancy are dynamic but accounting for these changes allows identification of an age-independent NK gene expression signature that tracks with rate of progression to T1D. Incorporating gene expression signatures alongside patterns of islet autoimmunity seroconversion facilitates robust prediction of individual risk, validated in an independent cohort. This creates the potential for early monitoring of at-risk infants for T1D onset, facilitating the prevention of severe complications such as ketoacidosis (*62*), effective trialling of preventive therapies or the identification of targets for immunomodulation (*63*).

Materials and Methods

Study design

TEDDY and nested case:control study design

Enrolment to the TEDDY study and design of the nested case-control biomarker discovery study is described in full elsewhere (*25*) and summarised here (fig. S2). In brief, the TEDDY study enrolled children <4.5 months of age from December 2004 to July 2010 through new-born screening for high-risk HLA-DR-DQ genotypes at six international centres (3 US, 3 EU). Written consent was obtained from primary carers for all participants, ethical approval was obtained from local institutional review boards and the study is monitored by an external evaluation committee formed by the National Institutes of Health. Blood samples were prospectively collected from 3 months of age, continuing at 3 monthly intervals until age 4, then every 6 months until age 15 unless seroconversion to persistent islet autoimmunity has occurred when they continued every 3 months until age 15. The primary endpoints of the TEDDY study are: (i) the appearance of persistent, confirmed islet autoimmunity, defined as the presence of one confirmed islet autoantibody (IAA, GAD65A or IA-2A) on at least two consecutive samples. islet autoimmunity result confirmation was obtained through reciprocal sample testing at two laboratories with the date of persistent seroconversion being the date of first detection of islet autoimmunity that was

subsequently shown to be persistent, and (ii) the clinical appearance of T1D, as defined by the American Diabetes Association diagnostic criteria (*64*).
Samples used for genomic analysis within the nested case:control study design used here were identified by risk set sampling in which islet autoimmunity and T1D controls were randomly selected from individuals who were free of the relevant event within 45 days of the case's event time using best available sample matching for clinical centre, sex, family history of T1D, and age (figS7). This identified two separate nested, matched cohorts each relating to one of the primary endpoints of the TEDDY study, namely T1D onset, and onset of islet autoimmunity (Fig. 1A, figure S2, table S2)(*25*).
The current study was designed to identify transcriptional coexpression networks in longitudinal whole blood transcriptomes in the TEDDY nested case:control study. Independent transcriptional networks were identified in and compared between subjects progressing to T1D or islet autoimmunity and age-matched controls (Fig S1) Eigengenes summarising coexpressed gene modules were then generated and modelled against the principal endpoints of the TEDDY study, namely the onset of islet autoimmunity and diagnosis of T1D. Association of early coexpression networks (measured in the earliest sampling timepoint for each individual) with later progression to either T1D or islet autoimmunity was also undertaken. For validation purposes, independent network analysis was undertaken of whole blood gene expression data from the publicly-available DIPP cohort (GSE30211).

RNA extraction and microarray hybridization
The TEDDY study collected 2.5 ml of peripheral blood to extract total RNA from enrolled children. Total RNA was extracted using a high throughput 96-well format extraction protocol using magnetic (MagMax) beads technology at the TEDDY RNA Laboratory, Jinfiniti Biosciences. Purified RNA (200 ng) was further used for cRNA amplification and labeling with biotin using Target Amp cDNA synthesis kit (Epicenter). Approximately 750ng of labeled cRNA was hybridized to the Illumina HumanHT-12 Expression BeadChips as per manufacturer's instructions. The HumanHT-12 Expression BeadChip provides coverage for more than 47,000 transcripts and known splice variants across the human transcriptome. After hybridization, arrays were washed, stained with Cy3-conjugated streptavidin, and scanned.

Microarray data preprocessing and normalization
The beadarray and lumi Bioconductor packages were used for preprocessing microarray data including image analysis, quality control, variance stabilization transformation, normalization and gene annotation. The MedianBackground method was used for local background correction. In addition, the BeadArray subversion of harshlight (BASH) method was used for beads artifact detection, which takes local spatial information into account when determining outliers. Each probe is replicated a varying number of times on each array; the summarization procedure produces a bead summary data in the form of a single signal intensity value for each probe. Illumina's default outlier function and modified mean and standard deviation were used to obtain a bead summary data. Variance-stabilizing transformation (vst) (*65*) and robust spline normalization (RSN) (*66*) method which combines the features of quantile and loess normalization were used for generating between-array normalization data. Quality control was performed by excluding arrays from further analysis with the corrupted image files, high gradient effects on the probe intensities, high percentage of beads

that were masked by the BASH method (*67*), low mean or median number of beads used to create the summary values for each probe on each array after outliers removal, low proportion of detected probes, low percentage of housekeeping genes expressed above the background level of the array, gender discrepancies using massiR package and poor pairwise array correlations. Transcriptional data from the DIPP cohort (GSE302011) was accessed from the NCBI-GEO repository using the GEOquery package from bioconductor in RStudio (version 3.5.1).

Transcriptomic QC and batch correction.
All the nested case-control pairs for the longitudinal transcriptome data were assigned to the same batch to constrain batchwise variation. In total, 2013 TEDDY samples were processed in 31 batches with a median batch size of 74 samples per batch (range: 18 to 86 samples per batch). In addition, two external QC samples (Donor 1 and Donor 2) were included in each batch to estimation of batch-to-batch variations. The MedianBackground method was used for local background correction. In addition, the BeadArray subversion of harshlight (BASH) method was used for beads artifact detection(*67*), which takes local spatial information into account when determining outliers. The first two principal components of the gene expression data before and after normalization, respectively are shown in figS8. The mean pairwise Pearson correlation coefficients after normalization were 0.97 (standard deviation (SD) = 0.04) for Donor 1 and 0.99 (SD = 0.01) for Donor 2.

Statistical analysis
Transcriptional network analysis
After data processing and quality control, 2013 samples from 401 individuals were included in the current analysis, representing 1698 samples from 342 individuals in the islet autoimmunity case:control study and 795 samples from 125 individuals in the T1D case:control study (fig. S2). For islet autoimmunity analyses, samples taken prior to onset of islet autoimmunity from both T1D and IA case:control cohorts were included along with their respective matched controls, stratified by the specificity of the first seroconversion as indicated. Transcriptional data was variance filtered (using the inflection point of cumulative median absolute deviation distribution) with data from 15,000 probes included in modular network analyses. The weighted gene coexpression networks (WGCNA) bioconductor package in Rstudio (version 3.5.1) was used to identify networks of co-expressed transcripts with scaled eigenvalues taken forward for lmm modelling. Scale-free topology was confirmed and a soft thresholding power selected by serial modelling of mean connectivity and adjacency functions. The network was constructed with a specified minimum module size of n=30 and medium sensitivity to cluster splitting (deepsplit =2). Independent networks were generated on cases and controls with comparison of network structure undertaken using WGCNA in Rstudio applying a composite preservation statistic as described(*68*) (fig. S1B). Modular structure in selected subgroups was visualised using tSNE plots using the Rtsne package from CRAN. As equivalent modular structure was identified in cases and controls, network analysis was repeated using the full cohort of 2013 samples to identify ‘universal’ modular eigengenes applicable to the entire cohort (rather than define them separately, fig. S1B).
For the DIPP cohort, the public dataset (GSE30211) was downloaded from GEO into R followed by filtering to retain unique genes, selecting those with the largest IQR per gene resulting in n=18469 features. This was mapped against the Refseq identifiers in

the TEDDY dataset to identify a matching set of n=9313 unique features that were used for modular network analyses as for the TEDDY dataset.

Longitudinal modelling

Longitudinal changes in gene expression were modelled by applying linear mixed modelling (lmm) to scaled modular eigenvalues using the lme4 package from CRAN in Rstudio (version 3.5.1). To identify changes in gene expression of cases that were not seen in matched controls, models were fitted for each modular eigenvector against either time to event (for cases, T1D diagnosis or islet autoimmunity onset) or to chronological age (matched controls) and the observed fit compared between cases and matched controls. Significance of effects was determined using a likelihood ratio test against a null model in the absence of that effect. This was repeated for additional covariates to test their independent association with progression rate including HLA subgroup, ethnicity and sex. For effects deemed significant (FDR<5%), specificity of association was determined by comparing observed significance in cases to that in controls in the form of a ratio of FDR values ($FDR_{T1D/IA}$:$FDR_{control}$). lmm were fitted including fixed terms (modular eigengene values and sex) and both random intercept and random slope terms for individuals. All identified modular signatures were iteratively tested with the extracted significance corrected for multiple testing using the Benjamini-Hochberg false discovery rate (FDR) method with a threshold for significance set at FDR 5%. Where indicated, for modular signatures of interest, modelling was repeated incorporating a natural cubic spline, implemented using the splines package in RStudio (version 3.5.1). Case:control lmm FDR ratios were visualised as radarcharts including all modules significantly associated with time to event, using the package radarchart from CRAN in Rstudio (version 3.5.1). Individual fits from lmm models were visualised using the ggplot2, sme and effects packages from CRAN in Rstudio (version 3.5.1).

Early and pre-T1D cohorts

For the TEDDY cohort, earliest available samples were identified from each individual and these were filtered for those obtained prior to IAb seroconversion (TEDDY preAb cohort, Fig4A). For the peri-T1D cohort, individuals were identified from whom a sample was taken within 365 days of diagnosis (serial IFN analysis) or the sample closest to diagnosis used (periT1D). The closest matched sample from the paired, matched control subject was used for comparative purposes. Gene expression modular signatures in these cross-sectional analyses were adjusted for sampling age, taking the residuals of a linear model including the relevant eigenvector and sampling age.

Module enrichment analysis

Module interpretation was performed using enrichment analysis against public repositories of defined transcriptional signatures as described in the text. Genes comprising selected modules were compared to reference signature repositories as indicated including ARCHS4, DICE and GO with a corrected Fishers exact test computed using Enrichr and visualised as the $-\log_{10}$ transformed adjusted value in a radar chart. Deconvolution analysis was undertaken using the CIBERSORT method against the LM22 dataset (*31*) with imputed cell proportions being correlated against module-specific eigenvectors. GPR171 was identified through a systematic screen against a relational database(*36*)) linking candidate 'druggable' targets (*35*) to associated transcriptional changes and other genomic data. T1D associated signatures

were screened against the existing IDG library in the ARCHS4 dataset comprising 352 'druggable' targets linked to 20883 genes. All targets showing any overlap with T1D signature genes were included in the radar plot visualisation (Fig2I) with only GPR171 achieving significant overlap.

NK cell analysis

Primary human NK cells obtained from healthy volunteers and stained with an excess of recombinantly engineered FcR-defective antibodies (CD3 clone REA641 and CD56 clone REA196, Miltenyi Biosciences) to avoid pre-activation. Flow sorting of NK cells ($CD3^{-}CD56^{+}$) was performed using an AriaIII sorter (BD) in the Cambridge BRC flow phenotyping hub. Purified NK cells were cultured for 48h in complete RPMI-1640 in the presence of target K562 cells and either a GPR171 inhibitor (MS21570, Tocris biotechne) or vehicle (PBS) and stained with an excess of antibodies against GPR171 (polyclonal rabbit anti-human GPR171, Abcam), CD71 (clone CY1G4, Biolegend) and Granzyme B (GZMB, clone GB11, Biolegend).

Bayesian joint modelling

For prediction we sought a method that could incorporate both baseline risk stratification and multiple longitudinal covariates to provide an estimate of event hazard with associated uncertainty. Joint models applied to longitudinal and survival data allow modelling of the error-free biomarker trajectories and disease process simultaneously and have several advantages over similar alternatives. Joint models have been shown to provide unbiased estimates of hazard ratios, unlike models using time-dependent covariates with increased performance compared to either baseline-only or time-dependent Cox models (*39, 69*). Joint modelling (jm) was performed using the mvjmbayes, jmbayes and coxph packages (*70*) from Bioconductor and CRAN in RStudio (version 3.5.1) to estimate the probability of getting disease at a given point in time given the data available. Let $S(t)$ denote the survival function, which we define to be $Pr(T_j^* > t)$, where T_j^* is the true time of getting disease for the jth patient. $S(t)$ is estimated using the hazard function $h(t)$, the instantaneous risk of getting disease

$$h(t) = \lim_{\delta t \to 0} \frac{Pr(t \leq T_j^* < t + \delta t | T_j^* \geq t)}{\delta t} \in [0, \infty)$$

Whereas standard approaches to survival analysis model and estimate the hazard function given the survival data at hand, Bayesian joint models allow hazard function estimation using both baseline covariates and longitudinal data in a proportional hazard model, using the predicted value from the lmm in the hazard model. This at once aims to reduce noise inherent in sparse biological data while not relying on assumptions that observations remain unchanged between measurements. The Bayesian methodology to compute the parameters allows for uncertainty estimates on predictions, achieved through Monte Carlo Markov chain sampling. We used the *JMbayes* package (*70*) from CRAN in RStudio (version 3.5.1).

The complete model including all covariates considered is given below:

$$\begin{aligned} h_i(t) = h_0(t) \exp\{ & \gamma Sex_i + \alpha_1(t) miaa\_sig_i + \alpha_2(t) gad\_sig'_i \\ & + \alpha_3(t) miaa\_ab_i + \alpha_4(t) gad\_ab_i + \alpha_4(t) ia2a\_ab_i \\ & + \alpha_6[miaa\_sig_i \times miaa\_ab_i] + \alpha_7[gad\_sig'_i \times gad\_ab_i] \\ & + \alpha_8[miaa\_ab_i \times gad\_ab_i] + \alpha_9[gad\_ab_i \times ia2a\_ab_i] \\ & + \alpha_{10}[ia2a\_ab_i \times miaa\_ab_i]\} \end{aligned}$$

Not all covariates were included in all model scenarios as an underlying goal was a sparse model incorporating covariates that can be measured in a simple, robust and cost-effective manner and which are likely to withstand later extension of the model into additional populations. Covariates used were factors that are known or suspected to correlate with progression to disease onset (IA) or progression (T1D diagnosis) and included whole blood transcriptional signatures and serial IAb data with time-varying effects (the hazard ratio was allowed to vary with time) and interaction effects between covariates (the type, number and sequence of IAb seroconversion was accounted for). Sex is included in our model as it has shown to correlate with T1D progression (*71*) (table S3) and is simple to obtain. HLA risk category is collected in the TEDDY study but was excluded to facilitate extension of the model between populations and ethnicities and because HLA risk groups also did not contribute to model performance on testing in the TEDDY discovery cohort (table S3).
Longitudinal data was fitted with a natural cubic spline-fitted lmm from the *JMbayes* library using the *mvglmer* function and survival predictions were made using *survFitJM* function. The input features to predict longitudinal outcome include the natural spline with 3 degrees of freedom fitted to time.

Predictive model performance estimates
When building any predictive model, it is imperative to balance predictive performance against the risk of 'overfitting', whereby the model performs well on a training dataset but fails to predict on unseen data. Predictive performance was first estimated using 10-fold cross-validation on the TEDDY discovery dataset.
Application in a clinical context was simulated by first making predictions on data collected up to 1 year of age, then serially increasing the amount of data available in steps of 0.15 years (mimicking clinical follow up), making disease predictions at each step over a constant time horizon of 1 year ahead. Model performance was evaluated using metrics addressing two key parameters, again using the JMbayes package: model discrimination (how well the model differentiates between individuals who do/do not reach an endpoint), and model calibration (how well the model predicts the observed data).
For discrimination, area under the receiver operator characteristic curve (AUC ROC) was selected to reflect both sensitivity and specificity of predictive accuracy. For calibration, prediction error (PE) was used as defined below. Each metric was applied to both cross-validated performance estimates on the discovery TEDDY cohort and following application of a 'fixed', optimal model from the discovery set to the independent validation DIPP cohort (which played no part in model training).

AUC is defined for a prediction horizon of Δt as follows

$$AUC(t, \Delta t) = Pr\Big[\pi_j(t+\Delta t|t) < \pi_{j'}(t+\Delta t|t) \;\Big|\; \{T_j^* \in (t, t+\Delta t]\} \cap \{T_{j'}^* > t+\Delta t\}\Big]$$

where $\pi_j(u|t)$ is the probability that patient j will survive up to time u given they are alive at time t and T_j^* is the true event time (T1D onset or IAb seroconversion).
Prediction error is defined as the expected loss given the difference between the predicted $N_i(u)$ and the true value $\pi_j(u|t)$ as given below:

$$PE(u|t) = E\big[L\{N_i(u) - \pi_j(u|t)\}\big]$$

Supplementary Materials

References and notes

1. J. A. Todd, Etiology of type 1 diabetes. *Immunity* **32**, 457-467 (2010); published online EpubApr 23 (10.1016/j.immuni.2010.04.001).
2. A. G. Ziegler, M. Rewers, O. Simell, T. Simell, J. Lempainen, A. Steck, C. Winkler, J. Ilonen, R. Veijola, M. Knip, E. Bonifacio, G. S. Eisenbarth, Seroconversion to multiple islet autoantibodies and risk of progression to diabetes in children. *JAMA* **309**, 2473-2479 (2013); published online EpubJun 19 (10.1001/jama.2013.6285).
3. M. A. Atkinson, M. von Herrath, A. C. Powers, M. Clare-Salzler, Current concepts on the pathogenesis of type 1 diabetes--considerations for attempts to prevent and reverse the disease. *Diabetes Care* **38**, 979-988 (2015); published online EpubJun (10.2337/dc15-0144).
4. S. A. Sharp, S. S. Rich, A. R. Wood, S. E. Jones, R. N. Beaumont, J. W. Harrison, D. A. Schneider, J. M. Locke, J. Tyrrell, M. N. Weedon, W. A. Hagopian, R. A. Oram, Development and Standardization of an Improved Type 1 Diabetes Genetic Risk Score for Use in Newborn Screening and Incident Diagnosis. *Diabetes Care* **42**, 200-207 (2019); published online EpubFeb (10.2337/dc18-1785).
5. A. G. Ziegler, G. T. Nepom, Prediction and pathogenesis in type 1 diabetes. *Immunity* **32**, 468-478 (2010); published online EpubApr 23 (10.1016/j.immuni.2010.03.018).
6. K. Cerosaletti, F. Barahmand-Pour-Whitman, J. Yang, H. A. DeBerg, M. J. Dufort, S. A. Murray, E. Israelsson, C. Speake, V. H. Gersuk, J. A. Eddy, H. Reijonen, C. J. Greenbaum, W. W. Kwok, E. Wambre, M. Prlic, R. Gottardo, G. T. Nepom, P. S. Linsley, Single-Cell RNA Sequencing Reveals Expanded Clones of Islet Antigen-Reactive CD4(+) T Cells in Peripheral Blood of Subjects with Type 1 Diabetes. *J Immunol* **199**, 323-335 (2017); published online EpubJul 1 (10.4049/jimmunol.1700172).
7. M. J. Dufort, C. J. Greenbaum, C. Speake, P. S. Linsley, Cell type-specific immune phenotypes predict loss of insulin secretion in new-onset type 1 diabetes. *JCI Insight* **4**, (2019); published online EpubFeb 21 (10.1172/jci.insight.125556).

8. P. N. Zakharov, H. Hu, X. Wan, E. R. Unanue, Single-cell RNA sequencing of murine islets shows high cellular complexity at all stages of autoimmune diabetes. *J Exp Med* **217**, (2020); published online EpubJun 1 (10.1084/jem.20192362).
9. N. Damond, S. Engler, V. R. T. Zanotelli, D. Schapiro, C. H. Wasserfall, I. Kusmartseva, H. S. Nick, F. Thorel, P. L. Herrera, M. A. Atkinson, B. Bodenmiller, A Map of Human Type 1 Diabetes Progression by Imaging Mass Cytometry. *Cell Metab* **29**, 755-768 e755 (2019); published online EpubMar 5 (10.1016/j.cmet.2018.11.014).
10. Q. Li, H. Parikh, M. D. Butterworth, A. Lernmark, W. Hagopian, M. Rewers, J. X. She, J. Toppari, A. G. Ziegler, B. Akolkar, O. Fiehn, S. Fan, J. P. Krischer, T. S. Group, Longitudinal Metabolome-Wide Signals Prior to the Appearance of a First Islet Autoantibody in Children Participating in the TEDDY Study. *Diabetes* **69**, 465-476 (2020); published online EpubMar (10.2337/db19-0756).
11. T. Vatanen, E. A. Franzosa, R. Schwager, S. Tripathi, T. D. Arthur, K. Vehik, A. Lernmark, W. A. Hagopian, M. J. Rewers, J. X. She, J. Toppari, A. G. Ziegler, B. Akolkar, J. P. Krischer, C. J. Stewart, N. J. Ajami, J. F. Petrosino, D. Gevers, H. Lahdesmaki, H. Vlamakis, C. Huttenhower, R. J. Xavier, The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* **562**, 589-594 (2018); published online EpubOct (10.1038/s41586-018-0620-2).
12. C. J. Stewart, N. J. Ajami, J. L. O'Brien, D. S. Hutchinson, D. P. Smith, M. C. Wong, M. C. Ross, R. E. Lloyd, H. Doddapaneni, G. A. Metcalf, D. Muzny, R. A. Gibbs, T. Vatanen, C. Huttenhower, R. J. Xavier, M. Rewers, W. Hagopian, J. Toppari, A. G. Ziegler, J. X. She, B. Akolkar, A. Lernmark, H. Hyoty, K. Vehik, J. P. Krischer, J. F. Petrosino, Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583-588 (2018); published online EpubOct (10.1038/s41586-018-0617-x).
13. U. Christen, M. G. von Herrath, Do viral infections protect from or enhance type 1 diabetes and how can we tell the difference? *Cell Mol Immunol* **8**, 193-198 (2011); published online EpubMay (10.1038/cmi.2010.71).
14. R. C. Ferreira, H. Guo, R. M. Coulson, D. J. Smyth, M. L. Pekalski, O. S. Burren, A. J. Cutler, J. D. Doecke, S. Flint, E. F. McKinney, P. A. Lyons, K. G. Smith, P. Achenbach, A. Beyerlein, D. B. Dunger, D. G. Clayton, L. S. Wicker, J. A. Todd, E. Bonifacio, C. Wallace, A. G. Ziegler, A type I interferon transcriptional signature precedes autoimmunity in children genetically at risk for type 1 diabetes. *Diabetes* **63**, 2538-2550 (2014); published online EpubJul (10.2337/db13-1777).
15. K. Vehik, K. F. Lynch, M. C. Wong, X. Tian, M. C. Ross, R. A. Gibbs, N. J. Ajami, J. F. Petrosino, M. Rewers, J. Toppari, A. G. Ziegler, J. X. She, A. Lernmark, B. Akolkar, W. A. Hagopian, D. A. Schatz, J. P. Krischer, H. Hyoty, R. E. Lloyd, T. S. Group, Prospective virome analyses in young children at increased genetic risk for type 1 diabetes. *Nat Med* **25**, 1865-1872 (2019); published online EpubDec (10.1038/s41591-019-0667-0).
16. F. Dotta, S. Censini, A. G. van Halteren, L. Marselli, M. Masini, S. Dionisi, F. Mosca, U. Boggi, A. O. Muda, S. Del Prato, J. F. Elliott, A. Covacci, R. Rappuoli, B. O. Roep, P. Marchetti, Coxsackie B4 virus infection of beta cells and natural killer cell insulitis in recent-onset type 1 diabetic

patients. *Proc Natl Acad Sci U S A* **104**, 5115-5120 (2007); published online EpubMar 20 (10.1073/pnas.0700442104).
17. A. M. Mehdi, E. E. Hamilton-Williams, A. Cristino, A. Ziegler, E. Bonifacio, K. A. Le Cao, M. Harris, R. Thomas, A peripheral blood transcriptomic signature predicts autoantibody development in infants at risk of type 1 diabetes. *JCI Insight* **3**, (2018); published online EpubMar 8 (10.1172/jci.insight.98212).
18. H. Kallionpaa, L. L. Elo, E. Laajala, J. Mykkanen, I. Ricano-Ponce, M. Vaarma, T. D. Laajala, H. Hyoty, J. Ilonen, R. Veijola, T. Simell, C. Wijmenga, M. Knip, H. Lahdesmaki, O. Simell, R. Lahesmaa, Innate immune activity is detected prior to seroconversion in children with HLA-conferred type 1 diabetes susceptibility. *Diabetes* **63**, 2402-2414 (2014); published online EpubJul (10.2337/db13-1775).
19. S. M. Cabrera, Y. G. Chen, W. A. Hagopian, M. J. Hessner, Blood-based signatures in type 1 diabetes. *Diabetologia* **59**, 414-425 (2016); published online EpubMar (10.1007/s00125-015-3843-x).
20. A. Olin, E. Henckel, Y. Chen, T. Lakshmikanth, C. Pou, J. Mikes, A. Gustafsson, A. K. Bernhardsson, C. Zhang, K. Bohlin, P. Brodin, Stereotypic Immune System Development in Newborn Children. *Cell* **174**, 1277-1292 e1214 (2018); published online EpubAug 23 (10.1016/j.cell.2018.06.045).
21. M. C. Arrieta, L. T. Stiemsma, P. A. Dimitriu, L. Thorson, S. Russell, S. Yurist-Doutsch, B. Kuzeljevic, M. J. Gold, H. M. Britton, D. L. Lefebvre, P. Subbarao, P. Mandhane, A. Becker, K. M. McNagny, M. R. Sears, T. Kollmann, C. S. Investigators, W. W. Mohn, S. E. Turvey, B. B. Finlay, Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci Transl Med* **7**, 307ra152 (2015); published online EpubSep 30 (10.1126/scitranslmed.aab2271).
22. E. F. McKinney, J. C. Lee, D. R. W. Jayne, P. A. Lyons, K. G. C. Smith, T-cell exhaustion, co-stimulation and clinical outcome in autoimmunity and infection. *Nature* **523**, 612-616 (2015); published online EpubJul 30 (10.1038/nature14468).
23. W. A. Hagopian, A. Lernmark, M. J. Rewers, O. G. Simell, J. X. She, A. G. Ziegler, J. P. Krischer, B. Akolkar, TEDDY--The Environmental Determinants of Diabetes in the Young: an observational clinical trial. *Ann N Y Acad Sci* **1079**, 320-326 (2006); published online EpubOct (10.1196/annals.1375.049).
24. E. Bonifacio, Predicting type 1 diabetes using biomarkers. *Diabetes Care* **38**, 989-996 (2015); published online EpubJun (10.2337/dc15-0101).
25. H. S. Lee, B. R. Burkhardt, W. McLeod, S. Smith, C. Eberhard, K. Lynch, D. Hadley, M. Rewers, O. Simell, J. X. She, B. Hagopian, A. Lernmark, B. Akolkar, A. G. Ziegler, J. P. Krischer, T. s. group, Biomarker discovery study design for type 1 diabetes in The Environmental Determinants of Diabetes in the Young (TEDDY) study. *Diabetes Metab Res Rev* **30**, 424-434 (2014); published online EpubJul (10.1002/dmrr.2510).
26. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008)1471-2105-9-559 [pii]
10.1186/1471-2105-9-559).

27. P. Langfelder, S. Horvath, Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* **1**, 54 (2007)10.1186/1752-0509-1-54).
28. J. M. Stuart, E. Segal, D. Koller, S. K. Kim, A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249-255 (2003); published online EpubOct 10 (10.1126/science.1087447).
29. T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, A. Soboleva, NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**, D991-995 (2013); published online EpubJan (10.1093/nar/gks1193).
30. A. Lachmann, D. Torre, A. B. Keenan, K. M. Jagodnik, H. J. Lee, L. Wang, M. C. Silverstein, A. Ma'ayan, Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun* **9**, 1366 (2018); published online EpubApr 10 (10.1038/s41467-018-03751-6).
31. A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, M. Diehn, A. A. Alizadeh, Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**, 773-782 (2019); published online EpubJul (10.1038/s41587-019-0114-2).
32. R. D. Leslie, S. T. Jerram, Stratifying Diabetes: Desperately Seeking Specificity. *Diabetes* **66**, 801-803 (2017); published online EpubApr (10.2337/dbi16-0074).
33. J. P. Krischer, K. F. Lynch, D. A. Schatz, J. Ilonen, A. Lernmark, W. A. Hagopian, M. J. Rewers, J. X. She, O. G. Simell, J. Toppari, A. G. Ziegler, B. Akolkar, E. Bonifacio, T. S. Group, The 6 year incidence of diabetes-associated autoantibodies in genetically at-risk children: the TEDDY study. *Diabetologia* **58**, 980-987 (2015); published online EpubMay (10.1007/s00125-015-3514-y).
34. B. J. Schmiedel, D. Singh, A. Madrigal, A. G. Valdovino-Gonzalez, B. M. White, J. Zapardiel-Gonzalo, B. Ha, G. Altay, J. A. Greenbaum, G. McVicker, G. Seumois, A. Rao, M. Kronenberg, B. Peters, P. Vijayanand, Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701-1715 e1716 (2018); published online EpubNov 29 (10.1016/j.cell.2018.10.022).
35. G. Rodgers, C. Austin, J. Anderson, A. Pawlyk, C. Colvis, R. Margolis, J. Baker, Glimmers in illuminating the druggable genome. *Nat Rev Drug Discov* **17**, 301-302 (2018); published online EpubMay (10.1038/nrd.2017.252).
36. A. D. Rouillard, G. W. Gundersen, N. F. Fernandez, Z. Wang, C. D. Monteiro, M. G. McDermott, A. Ma'ayan, The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database (Oxford)* **2016**, (2016)10.1093/database/baw100).
37. M. J. Haller, D. A. Schatz, The DIPP project: 20 years of discovery in type 1 diabetes. *Pediatr Diabetes* **17 Suppl 22**, 5-7 (2016); published online EpubJul (10.1111/pedi.12398).

38. J. H. Chen, S. M. Asch, Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *N Engl J Med* **376**, 2507-2509 (2017); published online EpubJun 29 (10.1056/NEJMp1702071).
39. D. Rizopoulos, *Joint models for longitudinal and time-to-event data : with applications in R*. Chapman & Hall/CRC biostatistics series (CRC Press, Boca Raton, 2012), pp. xiv, 261 pages.
40. D. Endesfelder, W. Zu Castell, E. Bonifacio, M. Rewers, W. A. Hagopian, J. X. She, A. Lernmark, J. Toppari, K. Vehik, A. J. K. Williams, L. Yu, B. Akolkar, J. P. Krischer, A. G. Ziegler, P. Achenbach, T. S. Group, Time-Resolved Autoantibody Profiling Facilitates Stratification of Preclinical Type 1 Diabetes in Children. *Diabetes* **68**, 119-130 (2019); published online EpubJan (10.2337/db18-0594).
41. M. Battaglia, S. Ahmed, M. S. Anderson, M. A. Atkinson, D. Becker, P. J. Bingley, E. Bosi, T. M. Brusko, L. A. DiMeglio, C. Evans-Molina, S. E. Gitelman, C. J. Greenbaum, P. A. Gottlieb, K. C. Herold, M. J. Hessner, M. Knip, L. Jacobsen, J. P. Krischer, S. A. Long, M. Lundgren, E. F. McKinney, N. G. Morgan, R. A. Oram, T. Pastinen, M. C. Peters, A. Petrelli, X. Qian, M. J. Redondo, B. O. Roep, D. Schatz, D. Skibinski, M. Peakman, Introducing the Endotype Concept to Address the Challenge of Disease Heterogeneity in Type 1 Diabetes. *Diabetes Care* **43**, 5-12 (2020); published online EpubJan (10.2337/dc19-0880).
42. M. Flodstrom-Tullberg, Y. T. Bryceson, F. D. Shi, P. Hoglund, H. G. Ljunggren, Natural killer cells in human autoimmunity. *Curr Opin Immunol* **21**, 634-640 (2009); published online EpubDec (10.1016/j.coi.2009.09.012).
43. M. Rodacki, B. Svoren, V. Butty, W. Besse, L. Laffel, C. Benoist, D. Mathis, Altered natural killer cells in type 1 diabetic patients. *Diabetes* **56**, 177-185 (2007); published online EpubJan (10.2337/db06-0493).
44. A. Oras, A. Peet, T. Giese, V. Tillmann, R. Uibo, A study of 51 subtypes of peripheral blood immune cells in newly diagnosed young type 1 diabetes patients. *Clin Exp Immunol* **198**, 57-70 (2019); published online EpubOct (10.1111/cei.13332).
45. L. Poirot, C. Benoist, D. Mathis, Natural killer cells distinguish innocuous and destructive forms of pancreatic islet autoimmunity. *Proc Natl Acad Sci U S A* **101**, 8102-8107 (2004); published online EpubMay 25 (10.1073/pnas.0402065101).
46. I. F. Lee, H. Qin, J. Trudeau, J. Dutz, R. Tan, Regulation of autoimmune diabetes by complete Freund's adjuvant is mediated by NK cells. *J Immunol* **172**, 937-942 (2004); published online EpubJan 15 (10.4049/jimmunol.172.2.937).
47. M. P. Nekoua, A. Bertin, F. Sane, E. K. Alidjinou, D. Lobert, J. Trauet, C. Hober, I. Engelmann, K. Moutairou, A. Yessoufou, D. Hober, Pancreatic beta cells persistently infected with coxsackievirus B4 are targets of NK cell-mediated cytolytic activity. *Cell Mol Life Sci* **77**, 179-194 (2020); published online EpubJan (10.1007/s00018-019-03168-4).
48. M. Flodstrom, A. Maday, D. Balakrishna, M. M. Cleary, A. Yoshimura, N. Sarvetnick, Target cell defense prevents the development of diabetes after viral infection. *Nat Immunol* **3**, 373-382 (2002); published online EpubApr (10.1038/ni771).

49. S. N. Waggoner, M. Cornberg, L. K. Selin, R. M. Welsh, Natural killer cells act as rheostats modulating antiviral T cells. *Nature* **481**, 394-398 (2012); published online EpubJan 19 (10.1038/nature10624).
50. I. Gomes, D. K. Aryal, J. H. Wardman, A. Gupta, K. Gagnidze, R. M. Rodriguiz, S. Kumar, W. C. Wetsel, J. E. Pintar, L. D. Fricker, L. A. Devi, GPR171 is a hypothalamic G protein-coupled receptor for BigLEN, a neuropeptide involved in feeding. *Proc Natl Acad Sci U S A* **110**, 16211-16216 (2013); published online EpubOct 1 (10.1073/pnas.1312938110).
51. J. F. Bach, L. Chatenoud, The hygiene hypothesis: an explanation for the increased frequency of insulin-dependent diabetes. *Cold Spring Harb Perspect Med* **2**, a007799 (2012); published online EpubFeb (10.1101/cshperspect.a007799).
52. U. Christen, T. Wolfe, U. Mohrle, A. C. Hughes, E. Rodrigo, E. A. Green, R. A. Flavell, M. G. von Herrath, A dual role for TNF-alpha in type 1 diabetes: islet-specific expression abrogates the ongoing autoimmune process when induced late but not early during pathogenesis. *J Immunol* **166**, 7023-7032 (2001); published online EpubJun 15 (10.4049/jimmunol.166.12.7023).
53. L. F. Lee, B. Xu, S. A. Michie, G. F. Beilhack, T. Warganich, S. Turley, H. O. McDevitt, The role of TNF-alpha in the pathogenesis of type 1 diabetes in the nonobese diabetic mouse: analysis of dendritic cell maturation. *Proc Natl Acad Sci U S A* **102**, 15995-16000 (2005); published online EpubNov 1 (10.1073/pnas.0508122102).
54. G. T. Nepom, M. Ehlers, T. Mandrup-Poulsen, Anti-cytokine therapies in T1D: Concepts and strategies. *Clin Immunol* **149**, 279-285 (2013); published online EpubDec (10.1016/j.clim.2013.02.003).
55. B. S. Marro, B. C. Ware, J. Zak, J. C. de la Torre, H. Rosen, M. B. Oldstone, Progression of type 1 diabetes from the prediabetic stage is controlled by interferon-alpha signaling. *Proc Natl Acad Sci U S A* **114**, 3708-3713 (2017); published online EpubApr 4 (10.1073/pnas.1700878114).
56. L. M. Duca, B. Wang, M. Rewers, A. Rewers, Diabetic Ketoacidosis at Diagnosis of Type 1 Diabetes Predicts Poor Long-term Glycemic Control. *Diabetes Care* **40**, 1249-1255 (2017); published online EpubSep (10.2337/dc17-0558).
57. L. A. Ferrat, K. Vehik, S. A. Sharp, A. Lernmark, M. J. Rewers, J. X. She, A. G. Ziegler, J. Toppari, B. Akolkar, J. P. Krischer, M. N. Weedon, R. A. Oram, W. A. Hagopian, T. S. Group, Committees, A combined risk score enhances prediction of type 1 diabetes among susceptible children. *Nat Med* **26**, 1247-1255 (2020); published online EpubAug (10.1038/s41591-020-0930-4).
58. E. Bonifacio, P. Achenbach, Birth and coming of age of islet autoantibodies. *Clin Exp Immunol* **198**, 294-305 (2019); published online EpubDec (10.1111/cei.13360).
59. K. Kang, Q. Meng, I. Shats, D. M. Umbach, M. Li, Y. Li, X. Li, L. Li, CDSeq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. *PLoS Comput Biol* **15**, e1007510 (2019); published online EpubDec (10.1371/journal.pcbi.1007510).
60. P. A. Lyons, M. Koukoulaki, A. Hatton, K. Doggett, H. B. Woffendin, A. N. Chaudhry, K. G. C. Smith, Microarray analysis of human leucocyte subsets:

the advantages of positive selection and rapid purification. *BMC Genomics* **8**, 64 (2007).
61. S. Aldridge, S. A. Teichmann, Single cell transcriptomics comes of age. *Nat Commun* **11**, 4307 (2020); published online EpubAug 27 (10.1038/s41467-020-18158-5).
62. A. M. Hekkala, J. Ilonen, J. Toppari, M. Knip, R. Veijola, Ketoacidosis at diagnosis of type 1 diabetes: Effect of prospective studies with newborn genetic screening and follow up of risk children. *Pediatr Diabetes* **19**, 314-319 (2018); published online EpubMar (10.1111/pedi.12541).
63. D. Bresson, M. von Herrath, Immunotherapy for the prevention and treatment of type 1 diabetes: optimizing the path from bench to bedside. *Diabetes Care* **32**, 1753-1768 (2009); published online EpubOct (10.2337/dc09-0373).
64. G. Puavilai, S. Chanprasertyotin, A. Sriphrapradaeng, Diagnostic criteria for diabetes mellitus and other categories of glucose intolerance: 1997 criteria by the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus (ADA), 1998 WHO consultation criteria, and 1985 WHO criteria. World Health Organization. *Diabetes Res Clin Pract* **44**, 21-26 (1999); published online EpubApr (
65. S. M. Lin, P. Du, W. Huber, W. A. Kibbe, Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res* **36**, e11 (2008); published online EpubFeb (10.1093/nar/gkm1075).
66. P. Du, W. A. Kibbe, S. M. Lin, lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547-1548 (2008); published online EpubJul 1 (10.1093/bioinformatics/btn224).
67. J. M. Cairns, M. J. Dunning, M. E. Ritchie, R. Russell, A. G. Lynch, BASH: a tool for managing BeadArray spatial artefacts. *Bioinformatics* **24**, 2921-2922 (2008); published online EpubDec 15 (10.1093/bioinformatics/btn557).
68. P. Langfelder, R. Luo, M. C. Oldham, S. Horvath, Is my network module preserved and reproducible? *PLoS Comput Biol* **7**, e1001057 (2011); published online EpubJan 20 (10.1371/journal.pcbi.1001057).
69. R. Henderson, P. Diggle, A. Dobson, Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465-480 (2000); published online EpubDec (10.1093/biostatistics/1.4.465).
70. D. Rizopoulos, The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC. *Journal of Statistical Software* **72**, (2016); published online Epub2016-08-28. (10.18637/jss.v072.i07).
71. J. P. Krischer, X. Liu, A. Lernmark, W. A. Hagopian, M. J. Rewers, J. X. She, J. Toppari, A. G. Ziegler, B. Akolkar, T. S. Group, The Influence of Type 1 Diabetes Genetic Susceptibility Regions, Age, Sex, and Family History on the Progression From Multiple Autoantibodies to Type 1 Diabetes: A TEDDY Study Report. *Diabetes* **66**, 3122-3129 (2017); published online EpubDec (10.2337/db17-0261).

Acknowledgements

We thank Dr. Angela Wood (Dept Public Health and Primary Care, University of Cambridge and Cambridge Centre for Aritificial Intelligence in Medicine, CCAIM) for helpful discussions and all patients and families for their participation.

Funding:
The TEDDY Study is funded by U01 DK63829, U01 DK63861, U01 DK63821, U01 DK63865, U01 DK63863, U01 DK63836, U01 DK63790, UC4 DK63829, UC4 DK63861, UC4 DK63821, UC4 DK63865, UC4 DK63863, UC4 DK63836, UC4 DK95300, UC4 DK100238, UC4 DK106955, UC4 DK112243, UC4 DK117483, and Contract No. HHSN267200700014C from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), National Institute of Environmental Health Sciences (NIEHS), Centers for Disease Control and Prevention (CDC), and JDRF. This work supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida (UL1 TR000064) and the University of Colorado (UL1 TR001082). KGCS is a Lister Prize fellow and is supported by a Wellcome Trust Senior Investigator award (200871/Z/16/Z). EFM is a Wellcome-Beit prize fellow (10406/Z/14/A) supported by the Wellcome Trust and Beit Foundation (10406/Z/14/Z) and by the National Institutes for Health Research Biomedical Research Centre (Cambridge). LPX's affiliation changed after completion of the manuscript and is now Département d'informatique et de recherche opérationnelle, Université de Montréal, Montréal, Canada and Mila, Quebec Institute for Learning Algorithms, Montréal, Canada.

Author Contributions
Clinical samples were collected and processed by the TEDDY study group as described. EFM analysed the data along and wrote the manuscript along with LPX with input from AL, EB, WAH, MJR, JXS, JT, KGCS, AGZ, BA, JPK and the TEDDY publications committee. OK undertook and analysed in vitro experiments along with EFM.

Competing Interests
AL is a member of the scientific advisory board of Diamyd Medical AB, AGZ is a member of the Data Safety and Monitoring Board of the Protect study for PreventionBio, JK provides advice to ActoBio, KGCS is a co-Founder of PredictImmune and Rheos Medicines and provides advice to Kymab, GSK and Elstar, EFM is a co-founder of PredictImmune and provides advice to Kymab.

Data and Materials Availability
The TEDDY study gene expression data that supports the findings of this study have been deposited in NCBI's database of Genotypes and Phenotypes (dbGaP) with the primary accession code phs001562.v1.p1.

Figure Captions

Figure 1. Dynamic changes in the infant blood transcriptome.
(A, B) Schematic illustration of the (A) TEDDY cohort (B) sampling from birth, through IAb seroconversion to T1D diagnosis illustrating population-level risFk of T1D. IAb+ samples may appear in both case:control cohorts hence subgroups do not

add up to the total. (C) tSNE plot illustrating the dissimilarity matrix of gene coexpression networks in T1D (left) and matched controls (right). Each dot represents a distinct gene (n=15000). Genes in both plots are coloured by modular assignment in the T1D coexpression network. (D) Scatterplot showing strength of T1D module preservation (y axis, $Z_{summary}$ score) in matched control data (red dashed line = strong preservation threshold, $Z_{summary} = 10$). (E) Line and scatterplots showing lmm effects (red, +/- 95% confidence interval [CI]) and gene expression eigenvalues (black dots) for 23 modular eigengenes showing significant (FDR<5%) age association in infancy. Colors are matched to Fig. 1 C-E. (F) Example module enrichment: line and scatterplot (left) showing lmm effects (red line, +/- 95% CI) for the 'yellow' module alongside radar plot (right) showing module enrichment (radial axis, $-\log_{10}$FDR) for cell type specific transcripts (G) Clustered heatmap illustrating significance ($-\log_{10}$FDR) of correlation (Pearson) of deconvolved cell subset proportions (y axis) against modular eigengene values (x axis). (H) Schematic line and scatterplot illustrating the use of lmm to compare modular gene expression signatures in matched cases and controls. (I) Radar plots showing all modules (arranged around plot circumference) associated with time to T1D onset (FDR<5%, left), association of the same modules with sampling age in matched controls (centre) and the ratio of observed significance in each (FDR_{T1D}:$FDR_{control}$, right). For radar plots, radial distance from the centre = $-\log_{10}$FDR, red line = threshold FDR <5%.

Figure 2. Age-independent changes in gene expression accompany T1D progression in subgroups defined by first IAb specificity.
(A, B) Radar plots showing age-corrected associations (lmm FDR_{T1D}:$FDR_{control}$) of all module eigengenes (shown around plot circumference) with time to T1D onset in IAA-first (A) and GADA-first (B) cases. Disease specificity indicated by a T1D:control significance ratio >1 ($-\log_{10}$FDR) = red line. (C, D) Line plots showing individual and summary (inset) effects of natural cubic spline-fitted lmm (+/-95%CI) fitted to IAAsig (C) or GADAsig (D) eigenvalues in cases (left) and matched controls (right). (E, F) Radar plot (E) and barplot (F) showing IAA module enrichment for cell type specific transcripts compared to ARCHS4 (E) and DICE (F) relational databanks. (G) Volcano plot (left) and line plots (right) showing all (left) and selected (right) correlations (Pearson) of IAAsig eigengenes against deconvoluted cell proportions. Significance threshold (FDR<5%) = red dashed line.
(H-J) Radar plot (H) and barplot (I) showing IAA module enrichment for cell type specific transcripts compared to ARCHS4 (H) and DICE (I) relational databanks. (J) Volcano plot (left) and line plots (right) showing all (left) and selected (right) correlations (Pearson) of GADAsig eigengenes against deconvoluted cell proportions. Significance threshold (FDR<5%) = red dashed line. For barplots (F, I) TPM = transcripts per million reads; expression (mean +/- sem) of IAAsig (F) and GADAsig (I) per cell type is shown. (K) Radar plot showing enrichment ($-\log_{10}$FDR) of IAAsig genes against 352 'druggable' targets linked to 20883 genes from the IDG repository. Drug targets showing any overlap with IAAsig genes are included around the radar plot circumference. (L) Barplot (mean +/-sem, n=3-12 per group) showing IAAsig eigengene expression and (M) representative histograms showing GPR171 surface protein expression in circulating immune cell subsets (GSE22886). (N) Representative contour plot (upper panel) and scatterplot (lower panel) showing GPR171 surface protein expression (y axis, $\log_{10}$ MFI) and Granzyme B (GZMB) protein expression (x axis) following coculture of purified primary human NK cells

with K562 target cells along with vehicle (left) or a titrated dose range of specific GPR171 inhibitor (Inh; right). * =P<0.05, MFI = median fluorescence index.

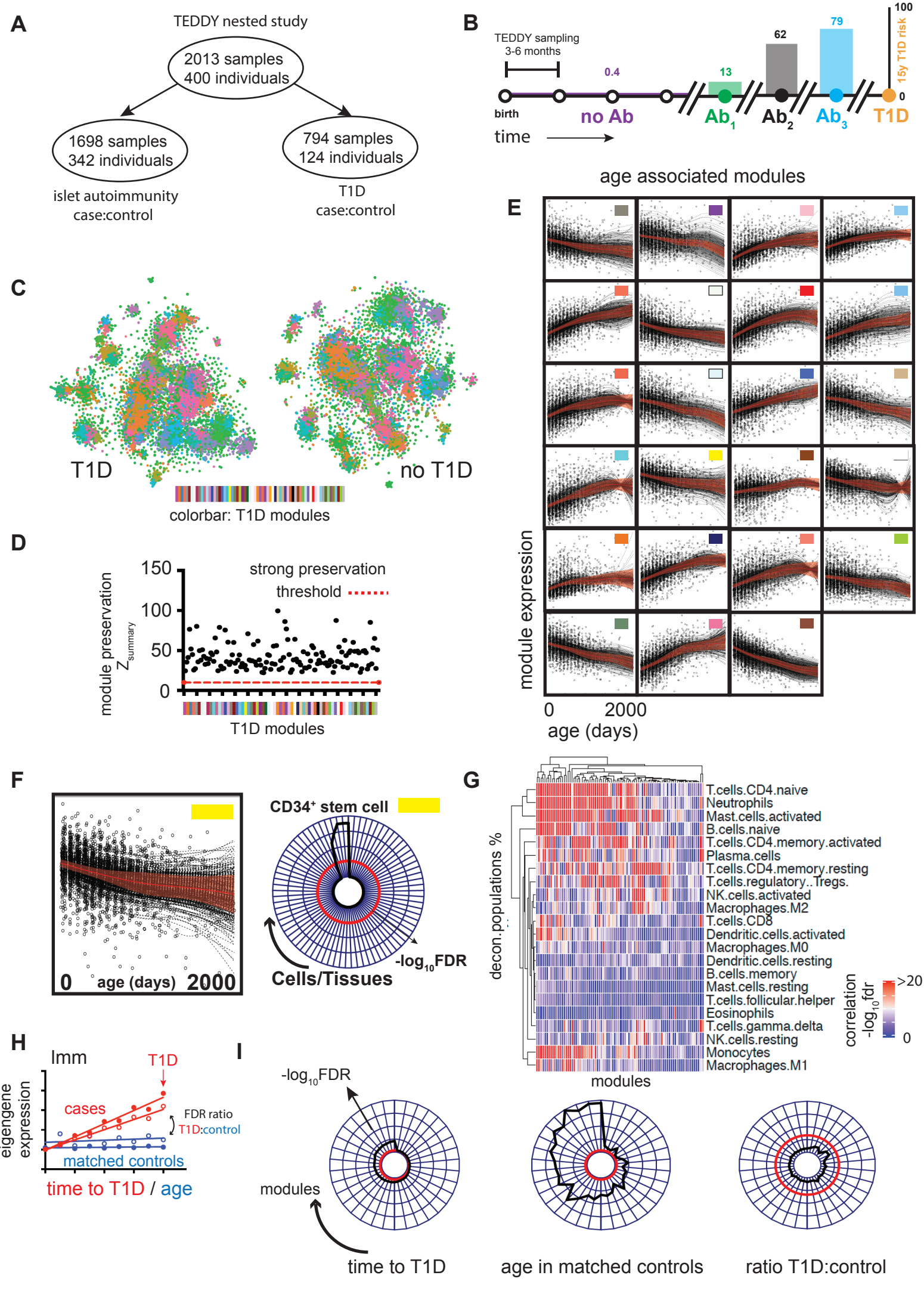
Figure 3. Age-independent changes in longitudinal gene expression accompany islet autoimmunity.
(A) Schematic line and scatterplot illustrating the use of a lmm to compare modular gene expression signatures in matched islet autoimmunity cases and healthy controls. (B, C) Radar plots showing age-independent associations (FDR_{IA}:$FDR_{control}$) of all gene expression modules significantly associated (FDR<5%) with time to islet autoimmunity onset in either IAA-first (B) or GADA-first (C) cases. Disease specificity indicated by a T1D:control significance ratio >1 ($-\log_{10}FDR$) = red line. (D) Line plots showing individual and summary (inset) effects of a natural cubic spline-fitted lmm (+/-95% CI) fitted to IAsig eigenvalues in IAA-first individuals (left, n= 82, vs time to islet autoimmunity) and matched controls (right, n= 63, vs age). (E) Line plots showing individual and summary (inset) effects of a natural cubic spline-fitted lmm (+/-95% CI) fitted to IAsig eigenvalues in GADA-first cases (left, n=54, v time to islet autiommunity) and matched controls (right, n=49, v age). (F) Radar plots showing IAsig module enrichment for cell type-specific transcripts, kinase targets, and transcription factor targets from the ARCHS4 dataset and barplot (right) showing cell specific expression of IAsig genes in the DICE dataset. TPM = transcripts per million reads, with expression (mean +/- sem) per cell type shown. (G) Line plots showing individual and summary (inset) effects of natural cubic spline-fitted lmm (+/-95% CI) fitted to the NK cell-enriched module eigenvalues (from Fig 2B) in IAA first cases (left, n=82, vs time to islet autoimmunity) and matched controls (right, n=63, vs age). (H) Line plots showing individual and summary (inset) effects of natural cubic spline-fitted lmm (+/-95% CI) fitted to the NK cell-enriched module eigenvalues (from Fig 2B) in GADA first cases (left, n=54, v time to islet autoimmunity) and matched controls (right, n=49, v age) (I, J) Line plots showing individual and summary (inset) effects of natural cubic spline-fitted lmm (+/-95% CI) fitted to the DIPP NK module in IA^{pos} cases (I, left, n=26, vs time to islet autoimmunity) and matched IA^{neg} controls (I, right, n=32, vs sampling age) and in T1D cases (J, left, n=24 vs time to T1D) and matched controls (J, right, n=34, vs sampling age). For radar plots (F, J) radial distance = $-\log_{10FDR}$ with threshold FDR (5%) in red.

Figure 4. Pre-seroconversion gene expression changes associate with rate of T1D progression.
(A) Schematic illustration and (B) density plot showing age distribution of earliest pre-seroconversion samples taken in the TEDDY transcriptomic study (TEDDY preAb). 'Outcome' indicates later progression rather than state at time of sampling. (C) Volcano plot showing correlation of all network modules (x axis, r) against adjusted significance ($-\log_{10}FDR$, y-axis). (D) Scatterplot illustrating inverse correlation of B lymphoblast (orange) and monocyte (black) modular signatures and their association with rate of T1D progression. (E) Radar plots illustrating enrichment of the orange, blue (upper), black and pink (lower) modules against the human cell atlas (left) and hallmark signature sets (right). Radial axis = $-\log_{10}FDR$, red line = significance threshold FDR (5%). (F) Scatterplot illustrating expression (mean+/-sem) of black, pink, orange, blue, and IFN response signatures by outcome group in the TEDDY preAb cohort, * = Mann Whitney P<0.05. Outcome group reflects final clinical status rather than status at the time of sampling. (G) Scatterplot showing age-

corrected eigengene expression of the 'black' monocyte/TNF-enriched signature in peri-T1D samples from TEDDY (n=54 samples within 12 months prior to diagnosis). (H) Scatterplot showing age-corrected eigengene expression of the 'black' monocyte/TNF-enriched signature in peri-T1D samples from the DIPP cohort (n=18 samples within 3 months prior to diagnosis) and their matched controls (n=18). (I) Radar plot showing modular enrichment for interferon (IFN) response transcripts. (J) Scatter and line plots showing lmm summary of longitudinal type1 IFN module expression (red line, +/- 95% CI shaded) in pre-T1D children (right, n=62) and matched controls (left, n=62). (K) Scatter and line plot illustrating 'spikes' of type 1 IFN response module and (L) age-matched peak expression in pre-T1D (red, n=57) and age-matched control samples (black, n=57) in the 12 months preceding T1D onset.

Figure 5. Validated prediction of T1D hazard in at-risk infants.
(A) Schematic of multivariate Bayesian joint model data input (top), model building (middle) and model performance assessment (bottom). (B) Illustration of cross-validation method employed on discovery TEDDY cohort. (C-F) Line and scatterplots showing predictive accuracy of models applied to each of two clinical scenarios (schematically illustrated above plots in red). Left, serial prediction over future 12-month horizon using cumulative data; right, serial prediction over extended future horizon using fixed data. (C) Model i incorporating IAb status ($IAb^{+/-}$) only. (D) model ii incorporating longitudinal IAb type, status, timing and interaction effects. (E) model iii incorporating model ii plus IAAsig/GADAsig and interaction effects. Predictive accuracy (AUC ROC) determined through 10-fold cross-validation on the discovery TEDDY dataset as illustrated in (B). (F) Predictive accuracy (AUC ROC) of independent validation of model iii on the DIPP dataset (G, H) Representative line and scatterplots illustrating serial prediction of individual T1D risk for a T1D case (G) and matched control (H) with predictions every 6 months, made over a horizon of 1 year. Error bars represent mean +/- SEM. Arrows indicate timing of seroconversion events.

Figure 1


A
TEDDY nested study
2013 samples
400 individuals
1698 samples
342 individuals
islet autoimmunity
case:control
794 samples
124 individuals
T1D
case:control
B
TEDDY sampling
3-6 months
0.4
13
62
79
100
15y T1D risk
0
birth
no Ab
Ab1
Ab2
Ab3
T1D
time
C
T1D
no T1D
colorbar: T1D modules
D
module preservation
Zsummary
150
100
50
0
strong preservation
threshold
T1D modules
E
age associated modules
module expression
0
2000
age (days)
F
0
age (days)
2000
CD34+ stem cell
-log10FDR
Cells/Tissues
G
decon.populations %
T.cells.CD4.naive
Neutrophils
Mast.cells.activated
B.cells.naive
T.cells.CD4.memory.activated
Plasma.cells
T.cells.CD4.memory.resting
T.cells.regulatory..Tregs.
NK.cells.activated
Macrophages.M2
T.cells.CD8
Dendritic.cells.activated
Macrophages.M0
Dendritic.cells.resting
B.cells.memory
Mast.cells.resting
T.cells.follicular.helper
Eosinophils
T.cells.gamma.delta
NK.cells.resting
Monocytes
Macrophages.M1
modules
correlation
-log10fdr
>20
0
H
lmm
T1D
cases
FDR ratio
T1D:control
matched controls
eigengene expression
time to T1D / age
I
-log10FDR
modules
time to T1D
age in matched controls
ratio T1D:control

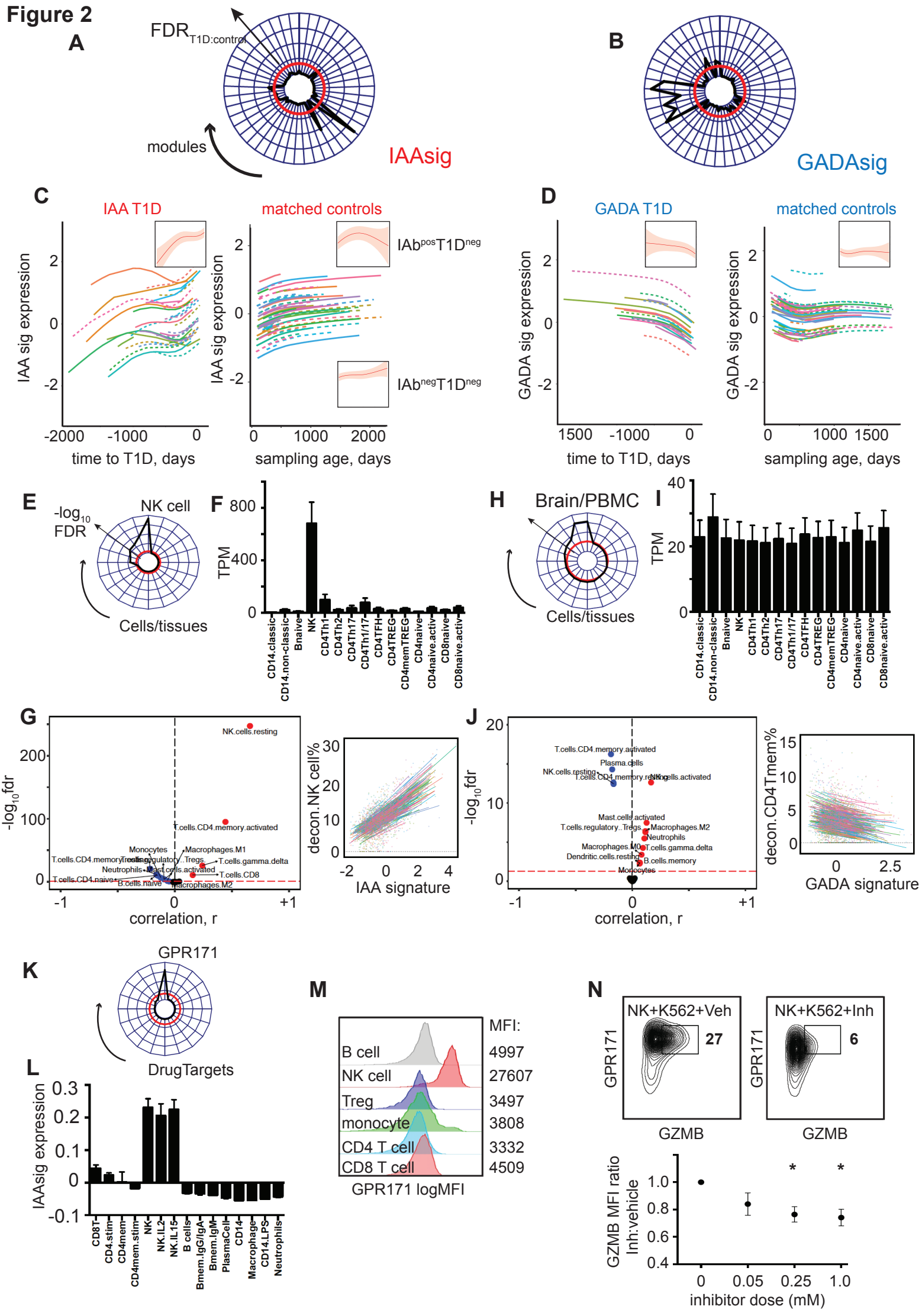


Figure 2
A
FDR T1D:control
modules
IAAsig
B
GADAsig
C
IAA T1D
matched controls
IAA sig expression
time to T1D, days
sampling age, days
IAbposT1Dneg
IAbnegT1Dneg
D
GADA T1D
matched controls
GADA sig expression
time to T1D, days
sampling age, days
E
NK cell
-log10 FDR
Cells/tissues
F
TPM
CD14.classic
CD14.non-classic
Bnaive
NK
CD4Th1
CD4Th2
CD4Th17
CD4Th1/17
CD4TFH
CD4TREG
CD4memTREG
CD4naive
CD4naive.activ
CD8naive
CD8naive.activ
H
Brain/PBMC
Cells/tissues
I
TPM
G
-log10fdr
correlation, r
NK.cells.resting
T.cells.CD4.memory.activated
Monocytes
Macrophages.M1
T.cells.CD4.memory.resting
T.cells.regulatory..Tregs
T.cells.gamma.delta
Neutrophils
Mast.cells.activated
T.cells.CD8
T.cells.CD4.naive
B.cells.naive
Macrophages.M2
decon.NK cell%
IAA signature
J
-log10fdr
correlation, r
T.cells.CD4.memory.activated
Plasma.cells
NK.cells.resting
NK.cells.activated
Mast.cells.activated
T.cells.regulatory..Tregs
Macrophages.M2
Neutrophils
Macrophages.M0
T.cells.gamma.delta
Dendritic.cells.resting
B.cells.memory
Monocytes
decon.CD4Tmem%
GADA signature
K
GPR171
DrugTargets
L
IAAsig expression
CD8T
CD4.stim
CD4mem
CD4mem.stim
NK
NK.IL2
NK.IL15
B cells
Bmem.IgG/IgA
Bmem.IgM
PlasmaCell
CD14
Macrophage
CD14.LPS
Neutrophils
M
MFI:
B cell 4997
NK cell 27607
Treg 3497
monocyte 3808
CD4 T cell 3332
CD8 T cell 4509
GPR171 logMFI
N
NK+K562+Veh
NK+K562+Inh
GPR171
GZMB
27
6
GZMB MFI ratio Inh:vehicle
inhibitor dose (mM)

Figure 3

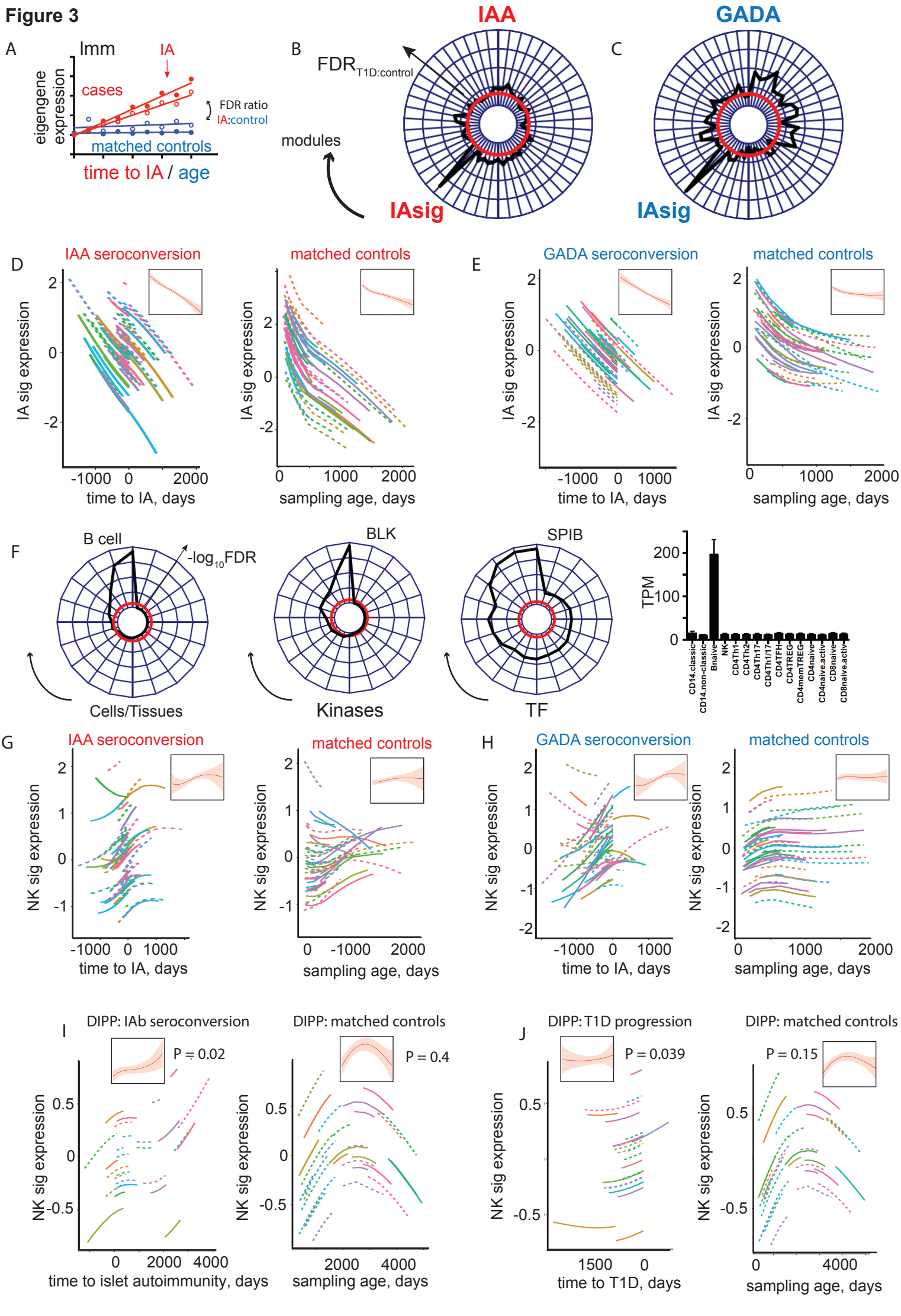

A
lmm
IA
cases
matched controls
eigengene expression
FDR ratio IA:control
time to IA / age
B
IAA
FDR T1D:control
modules
IAsig
C
GADA
IAsig
D
IAA seroconversion
matched controls
IA sig expression
time to IA, days
sampling age, days
E
GADA seroconversion
matched controls
IA sig expression
time to IA, days
sampling age, days
F
B cell
-log10FDR
Cells/Tissues
BLK
Kinases
SPIB
TF
TPM
CD14.classic
CD14.non-classic
Bnaive
NK
CD4Th1
CD4Th2
CD4Th17
CD4Th1/17
CD4TFH
CD4TREG
CD4memTREG
CD4naive
CD4naive.activ
CD8naive
CD8naive.activ
G
IAA seroconversion
matched controls
NK sig expression
time to IA, days
sampling age, days
H
GADA seroconversion
matched controls
NK sig expression
time to IA, days
sampling age, days
I
DIPP: IAb seroconversion
DIPP: matched controls
P = 0.02
P = 0.4
NK sig expression
time to islet autoimmunity, days
sampling age, days
J
DIPP: T1D progression
DIPP: matched controls
P = 0.039
P = 0.15
NK sig expression
time to T1D, days
sampling age, days

Figure 4

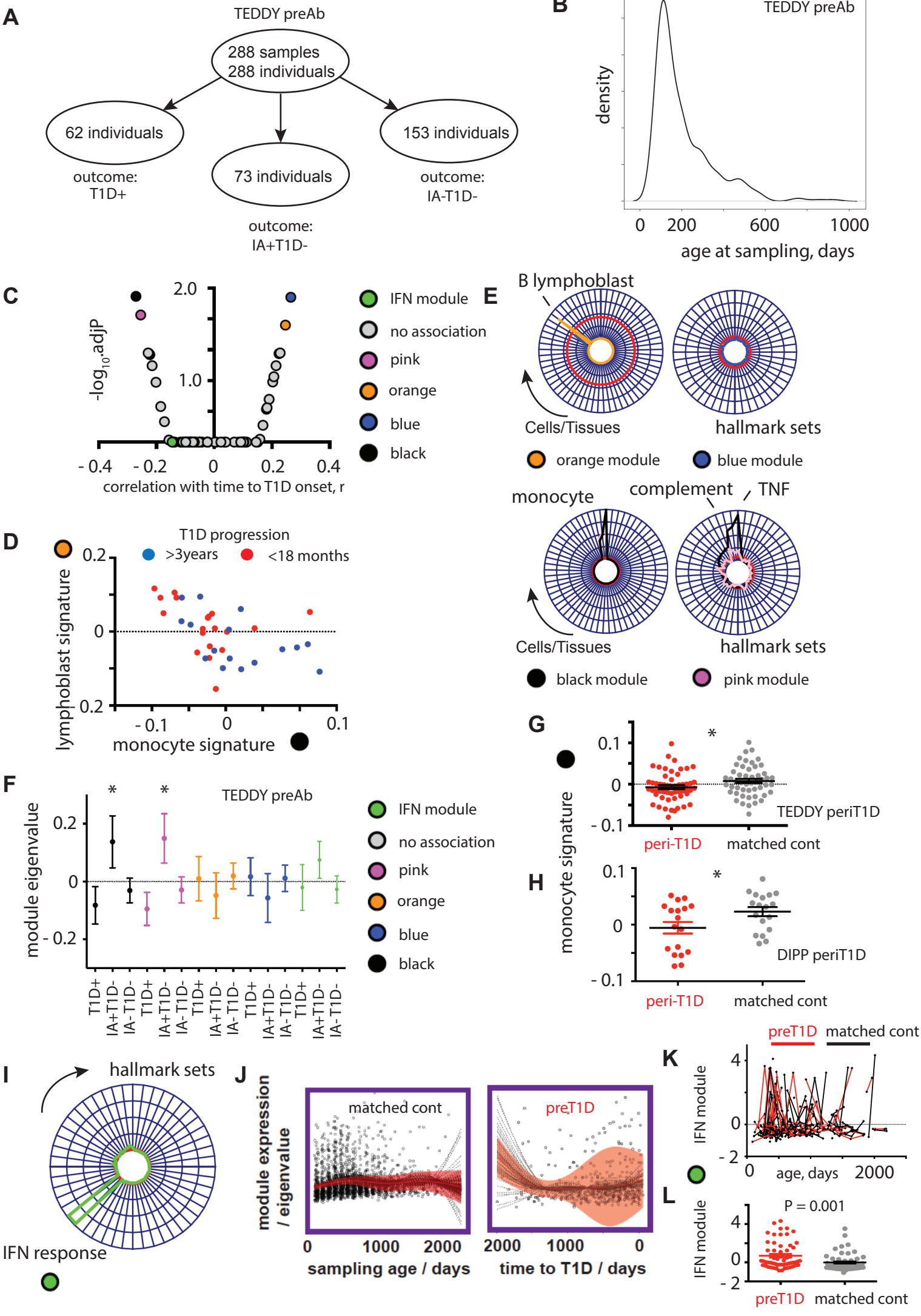
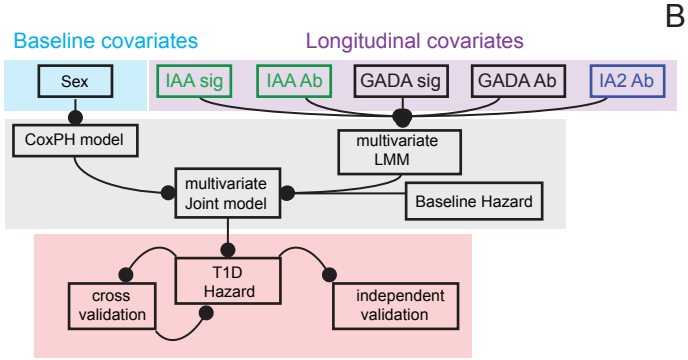
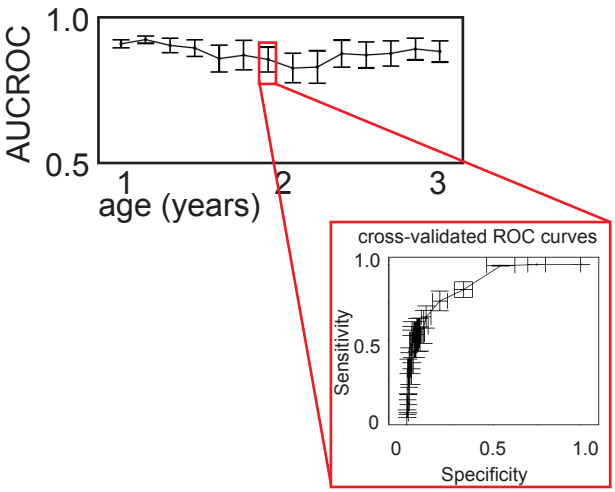
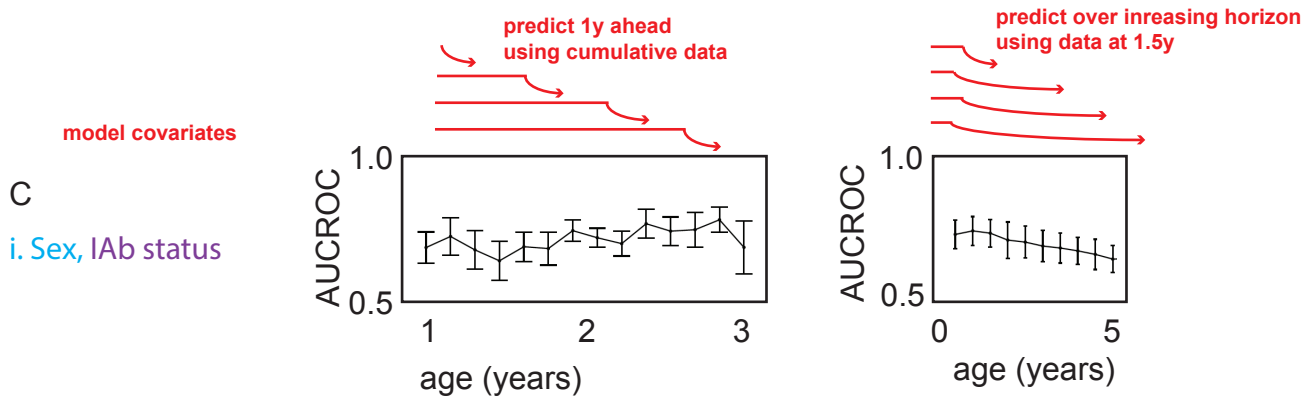

A
TEDDY preAb
288 samples
288 individuals
62 individuals
outcome:
T1D+
73 individuals
outcome:
IA+T1D-
153 individuals
outcome:
IA-T1D-
B
TEDDY preAb
density
0
200
600
1000
age at sampling, days
C
2.0
1.0
-log10.adjP
- 0.4
- 0.2
0
0.2
0.4
correlation with time to T1D onset, r
IFN module
no association
pink
orange
blue
black
D
T1D progression
>3years
<18 months
0.2
0
0.2
lymphoblast signature
- 0.1
0
0.1
monocyte signature
E
B lymphoblast
Cells/Tissues
hallmark sets
orange module
blue module
monocyte
complement
TNF
Cells/Tissues
hallmark sets
black module
pink module
F
TEDDY preAb
module eigenvalue
0.2
0
- 0.2
T1D+
IA+T1D-
IA-T1D-
IFN module
no association
pink
orange
blue
black
G
0.1
0
- 0.1
monocyte signature
TEDDY periT1D
peri-T1D
matched cont
H
DIPP periT1D
peri-T1D
matched cont
I
hallmark sets
IFN response
J
module expression / eigenvalue
matched cont
preT1D
0
1000
2000
sampling age / days
2000
1000
0
time to T1D / days
K
preT1D
matched cont
IFN module
4
0
- 2
0
age, days
2000
L
IFN module
P = 0.001
4
0
- 2
preT1D
matched cont

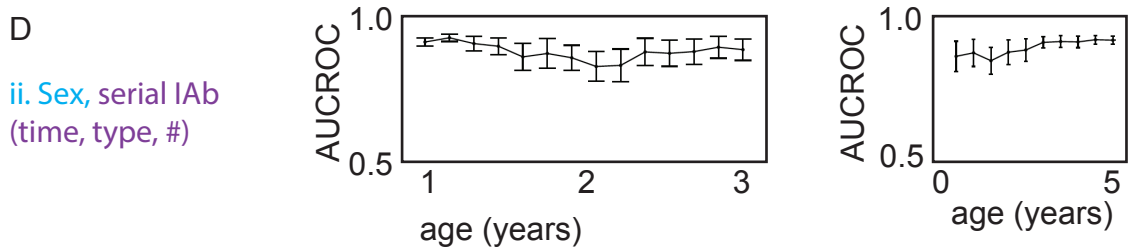
Figure 5 A

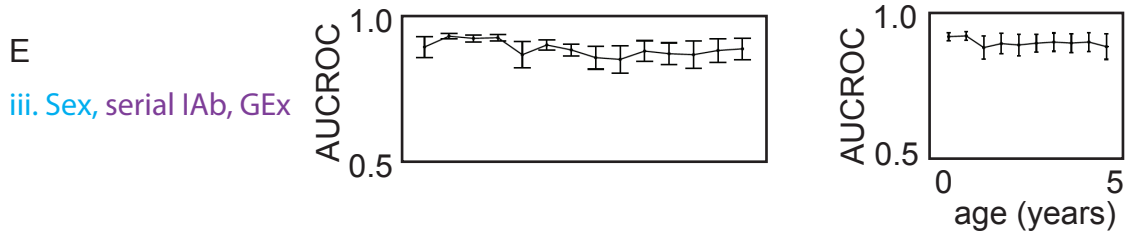

Baseline covariates
Longitudinal covariates
Data
Sex
IAA sig
IAA Ab
GADA sig
GADA Ab
IA2 Ab
Model building
CoxPH model
multivariate LMM
multivariate Joint model
Baseline Hazard
Prediction and Performance
T1D Hazard
cross validation
independent validation

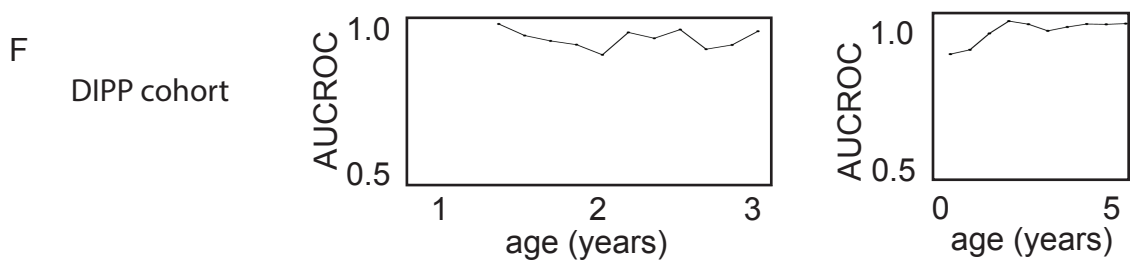
B


AUCROC
1.0
0.5
1
2
3
age (years)
cross-validated ROC curves
Sensitivity
Specificity
0
0.5
1.0


predict 1y ahead using cumulative data
predict over inreasing horizon using data at 1.5y
model covariates
C
i. Sex, IAb status
AUCROC
age (years)


D
ii. Sex, serial IAb (time, type, #)
AUCROC
age (years)


E
iii. Sex, serial IAb, GEx
AUCROC
age (years)


F
DIPP cohort
AUCROC
age (years)

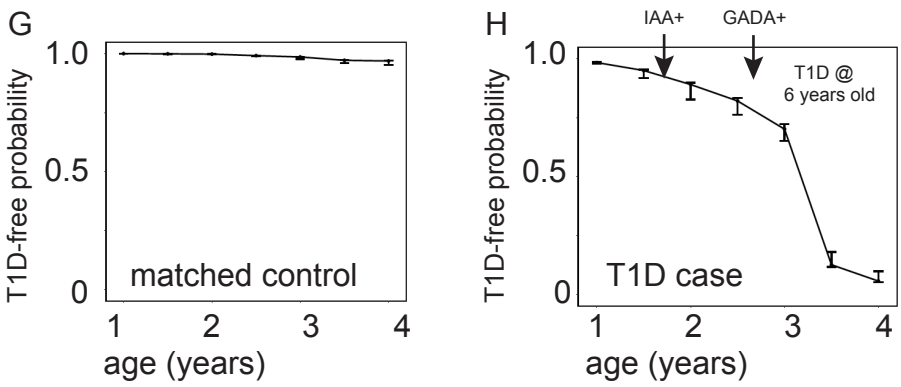
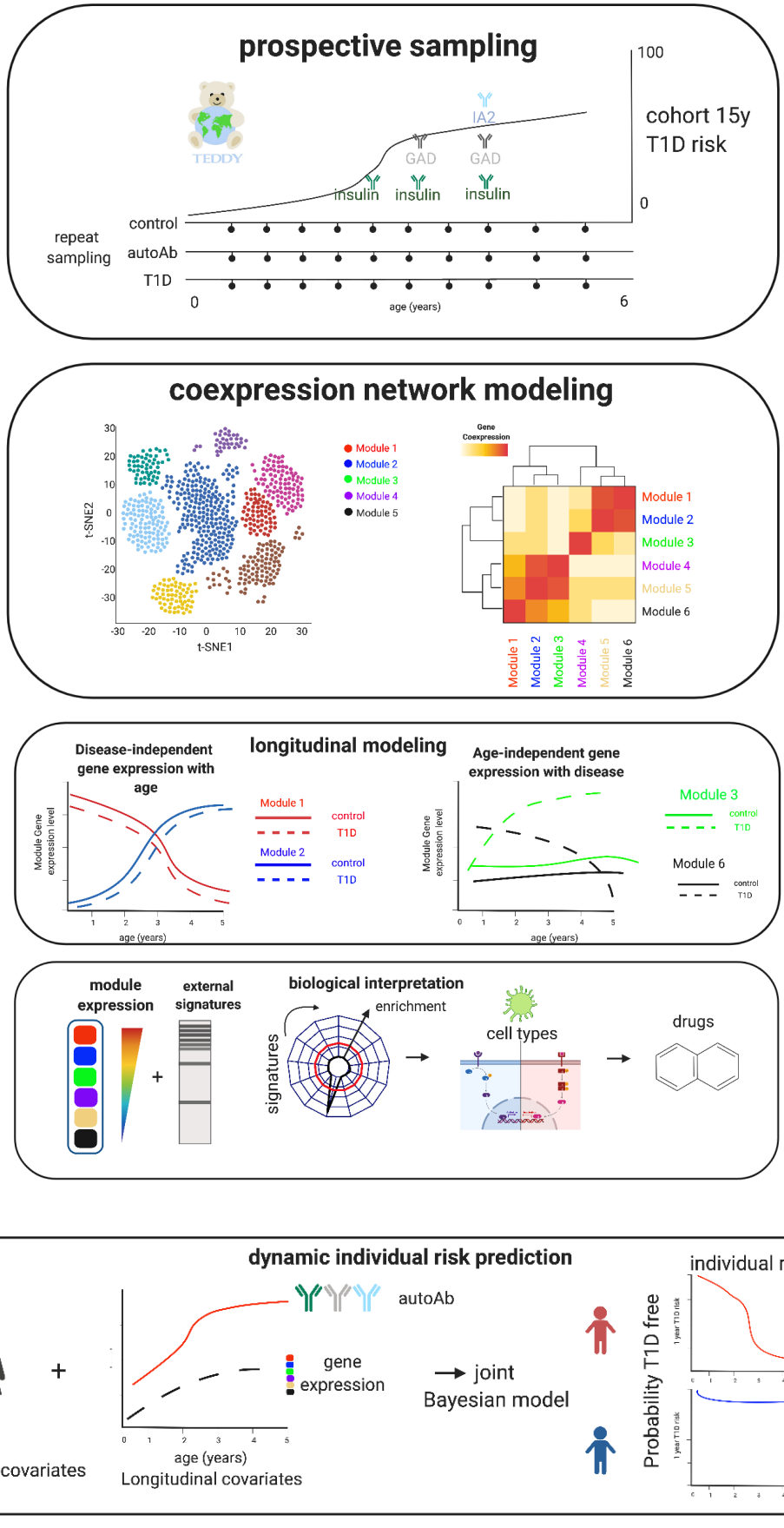

G
T1D-free probability
matched control
age (years)
H
IAA+
GADA+
T1D @ 6 years old
T1D case
age (years)

Fig S1

A


prospective sampling
TEDDY
IA2
GAD
GAD
insulin
insulin
insulin
100
cohort 15y
T1D risk
0
repeat
sampling
control
autoAb
T1D
0
6
age (years)
coexpression network modeling
t-SNE2
t-SNE1
Module 1
Module 2
Module 3
Module 4
Module 5
Module 6
Gene
Coexpression
longitudinal modeling
Disease-independent
gene expression with
age
Module Gene
expression level
age (years)
control
T1D
Age-independent gene
expression with disease
module
expression
external
signatures
biological interpretation
enrichment
signatures
cell types
drugs
dynamic individual risk prediction
autoAb
gene
expression
joint
Bayesian model
Baseline covariates
Longitudinal covariates
individual risk
Probability T1D free
T1D
no
T1D

B

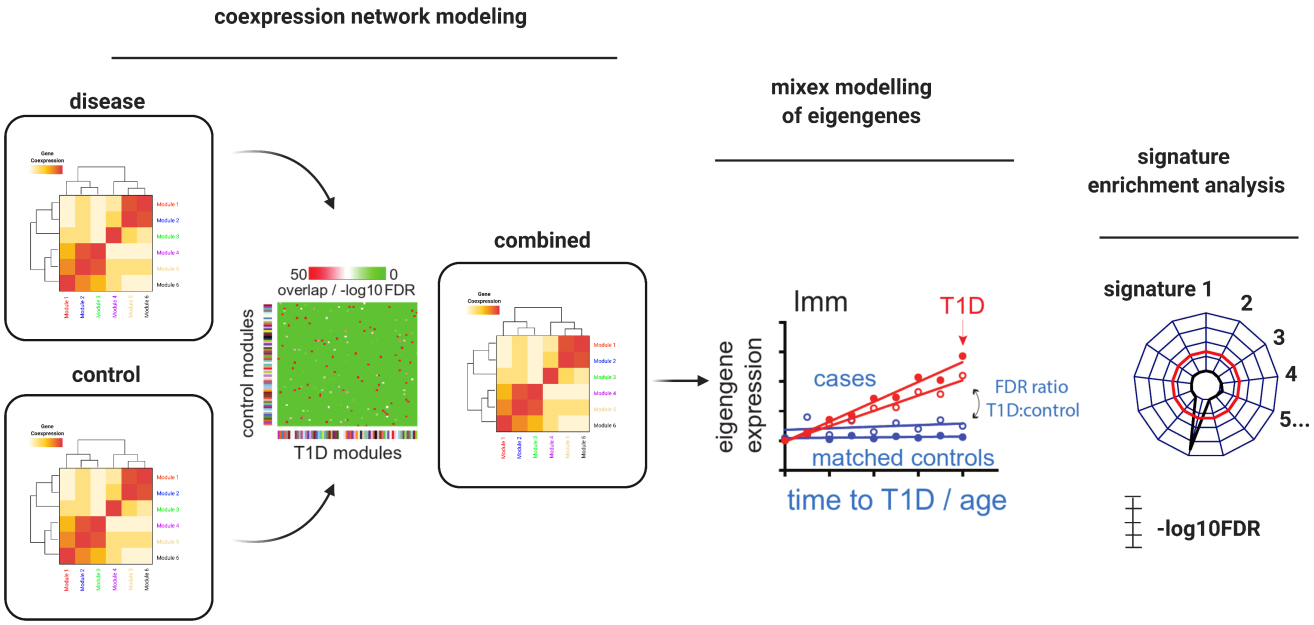
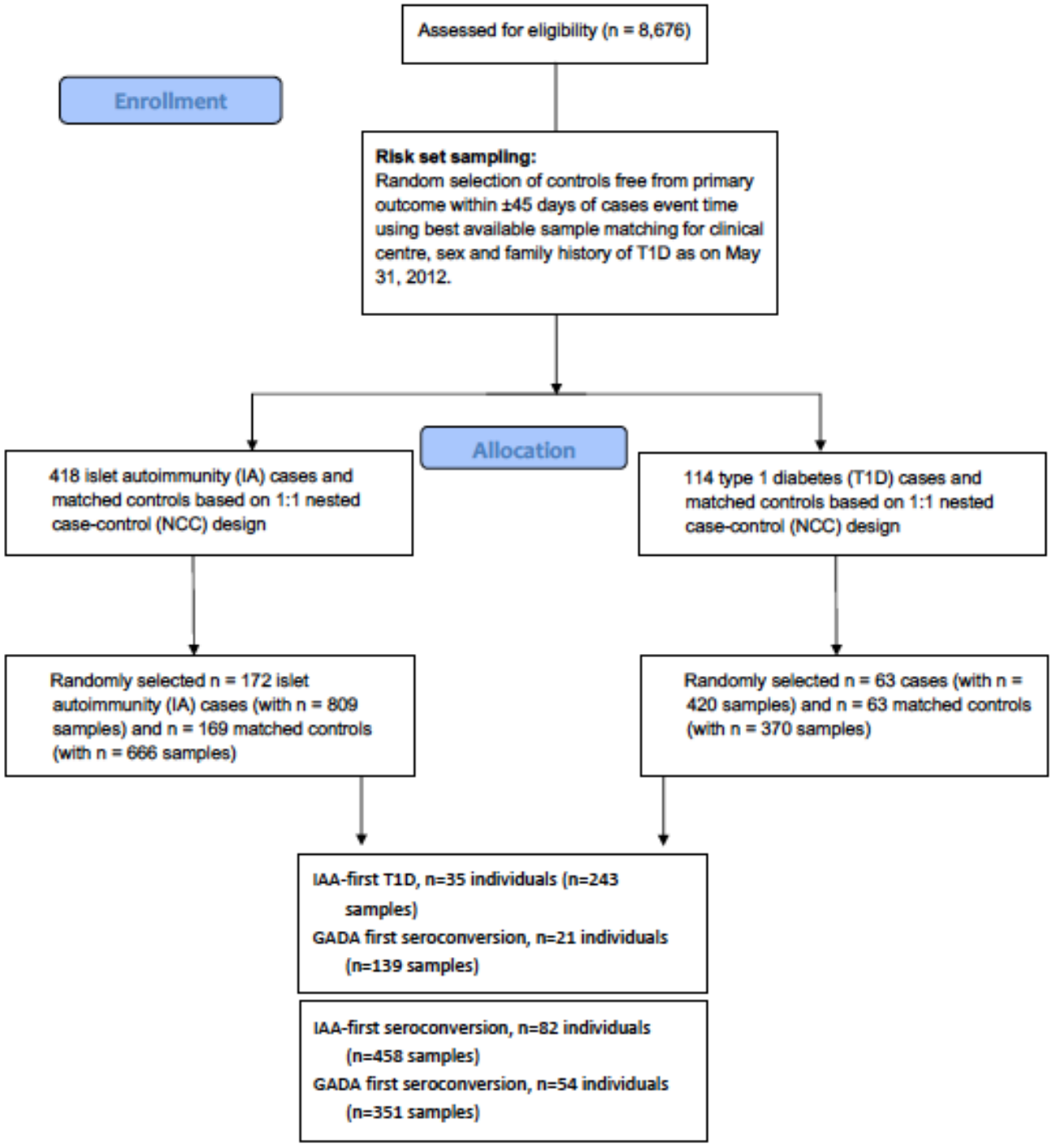


Figure S1

(A) Graphical abstract and analysis overview. **(A)** Graphical abstract illustrating the stages of the current analysis. (Top) Prospective sample collection in the TEDDY study, with samples taken every 3-6 months towards IA, T1D or neither in at risk children. (Second from top) Coexpression network modelling of peripheral whole blood samples to identify modules of coexpressed genes and comparative preservation in different states. (Middle) Longitudinal mixed modelling of summary eigenvalues from each coexpression module to identify age-associated changes in infancy and disease-specific changes relative to either T1D progression or IA onset. (Second from bottom) Biological interpretation of gene coexpression modules, identifying enrichment of external cell, transcription factor, kinase and drug response signatures within disease-specific coexpression modules. (Bottom) Building a predictive model using a combination of baseline (sex) and longitudinal (IAb, gene expression) covariates to facilitate dynamic, individual risk prediction over a near-term horizon. **(B)** Schematic illustration of the network coexpression analysis workflow in the TEDDY cohort. Network analysis was performed on T1D and matched control sets independently (left) with formal comparison indicating comparable modular structure in each (heatmap). A combined network was therefore used to identify modular eigengenes of coexpressed genes ('combined'). Comparison of linear mixed models incorporating longitudinal changes in eigengene signatures in disease (T1D or islet autoimmunity) compared to age-matched controls facilitated identification of disease-specific changes. Interpretation of disease-specific modules was undertaken using enrichment analysis.

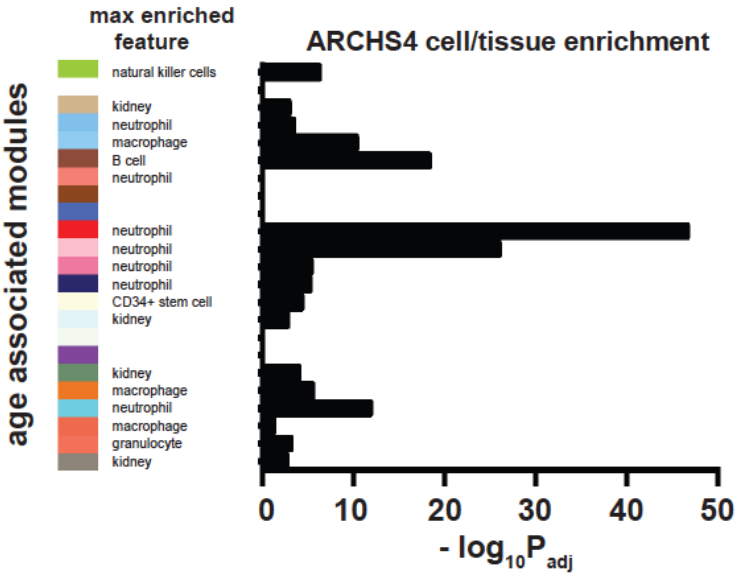
FigS2: TEDDY nested case:control studies CONSORT flowchart



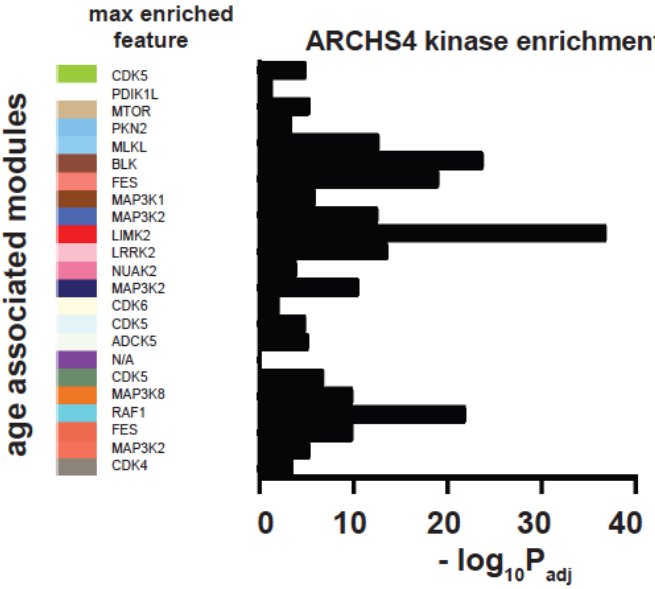
Supplementary Figure 2
Schematic illustration of the nested case:control transcriptomic study established within TEDDY. Further description available in Ref #14. Note that IAb^{+} samples may appear in both case:control cohorts and some T1D cases had either none (IAb^{-}) or multiple ($IAA^{+}GADA^{+}$) IAb during progression, hence subgroups may not add up to the totals shown.

Fig.S3

A


max enriched feature
ARCHS4 cell/tissue enrichment
age associated modules
natural killer cells
kidney
neutrophil
macrophage
B cell
neutrophil
neutrophil
neutrophil
neutrophil
neutrophil
CD34+ stem cell
kidney
kidney
macrophage
neutrophil
macrophage
granulocyte
kidney
0
10
20
30
40
50
- log10Padj

B


max enriched feature
ARCHS4 kinase enrichment
age associated modules
CDK5
PDIK1L
MTOR
PKN2
MLKL
BLK
FES
MAP3K1
MAP3K2
LIMK2
LRRK2
NUAK2
MAP3K2
CDK6
CDK5
ADCK5
N/A
CDK5
MAP3K8
RAF1
FES
MAP3K2
CDK4
0
10
20
30
40
- log10Padj

C

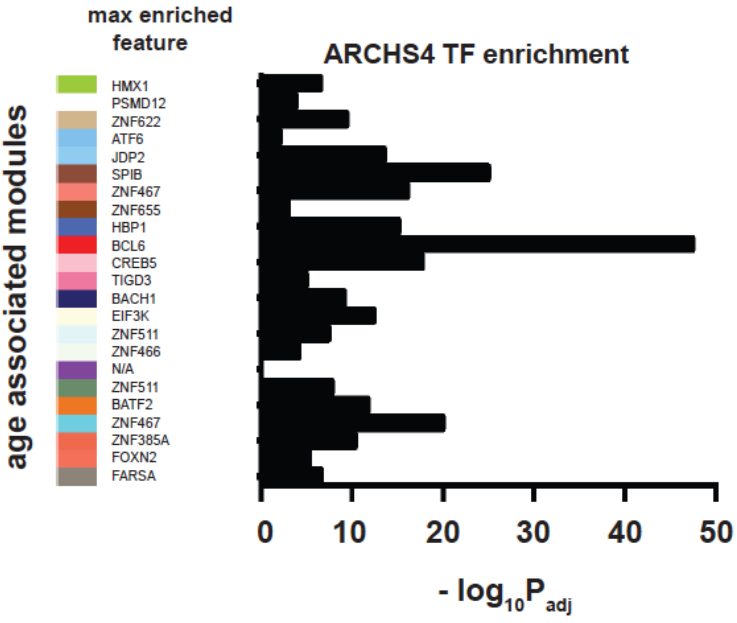

max enriched feature
ARCHS4 TF enrichment
age associated modules
HMX1
PSMD12
ZNF622
ATF6
JDP2
SPIB
ZNF467
ZNF655
HBP1
BCL6
CREB5
TIGD3
BACH1
EIF3K
ZNF511
ZNF466
N/A
ZNF511
BATF2
ZNF467
ZNF385A
FOXN2
FARSA
0
10
20
30
40
50
- log10Padj

Figure S3. Age-associated module enrichment. (A-C) Histogram plots showing maximal enrichment (x-axis, -log10FDR) of all gene expression modules associated with age in healthy infancy (colored blocks) for cell/tissue-specific transcripts **(A)**, transcription factor (TF) targets **(B)** and kinase targets **(C)**. In each case the most significant enrichment is shown along with its associated significance. Note that tissue-specific transcripts are included in the relational database, consequently 'kidney' expression will include infiltrating immune populations and is not restricted to kidney cells alone.

Fig. S4

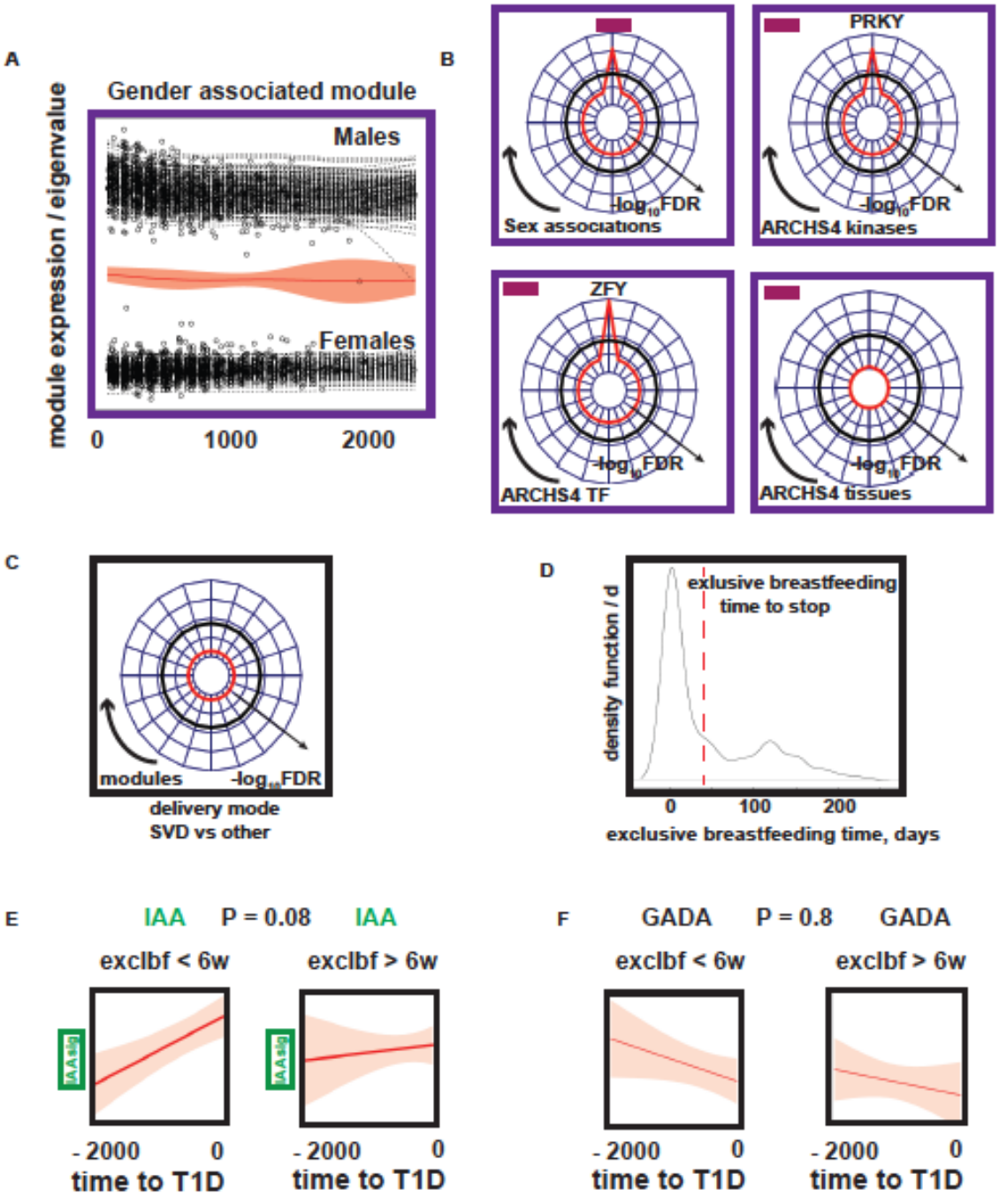

A
Gender associated module
Males
Females
module expression / eigenvalue
0
1000
2000
B
Sex associations
PRKY
ARCHS4 kinases
ZFY
ARCHS4 TF
ARCHS4 tissues
-log10FDR
C
modules
-log10FDR
delivery mode
SVD vs other
D
exlusive breastfeeding
time to stop
density function / d
0
100
200
exclusive breastfeeding time, days
E
IAA
P = 0.08
IAA
excIbf < 6w
excIbf > 6w
IAAstg
- 2000
0
time to T1D
F
GADA
P = 0.8
GADA
excIbf < 6w
excIbf > 6w
- 2000
0
time to T1D

Figure S4. Sex and breast-feeding associations of modular expression changes. (A) Line and scatterplot showing longitudinal changes in expression of the single gene expression module associated with sex (B, top left) along with linear mixed model effects (red line, +/- 95% CI). **(B)** Radar plots showing significance ($-\log_{10}$FDR, radial axis) of association between gene expression modular eigengenes and sex (top left), enrichment for kinases (top right), transcription factor (TF) targets (bottom left) and cell/tissue-specific transcripts (bottom right) in the ARCHS4 repository. **(C)** Radar plot showing significance (likelihood ratio test, $-\log_{10}$FDR, radial axis) of association between modular eigengene signatures and delivery mode (spontaneous vaginal delivery (SVD) v other, including caesarean section or instrumental delivery). **(D)** Density plot showing bimodal distribution of time to stopping exclusive breastfeeding in the full cohort (n=401 individuals). Red line indicates a 6-week threshold defining subgroups considered in panels E, F. **(E, F)** Model effects plots (+/-95% CI) summarising longitudinal expression of IAAsig (**E**, green) and GADAsig (**F**, black) for individuals with <6 weeks (**E, F** left) and > 6 weeks exclusive breastfeeding (**E, F** right). P = Likelihood ratio test.

Fig. S5

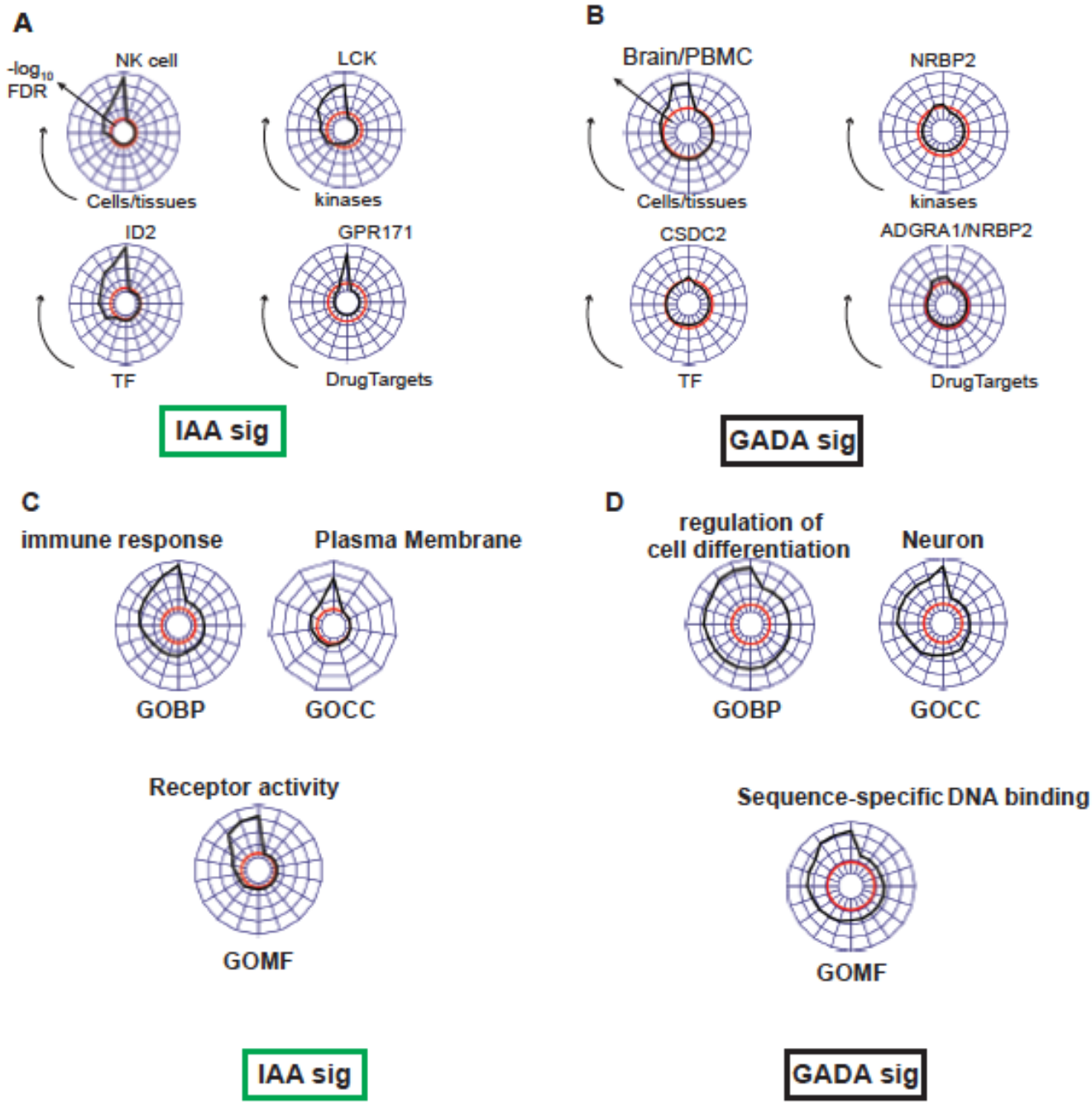

A
-log10 FDR
NK cell
Cells/tissues
LCK
kinases
ID2
TF
GPR171
DrugTargets
IAA sig
B
Brain/PBMC
Cells/tissues
NRBP2
kinases
CSDC2
TF
ADGRA1/NRBP2
DrugTargets
GADA sig
C
immune response
GOBP
Plasma Membrane
GOCC
Receptor activity
GOMF
IAA sig
D
regulation of cell differentiation
GOBP
Neuron
GOCC
Sequence-specific DNA binding
GOMF
GADA sig

Figure S5. IAA and GADA signature enrichment. Radarplots showing enrichment of the TEDDY IAA and GADA signatures against the ARCHS4 **(A, B)** databank for cell types (top left), kinases (top right), transcription factors (bottom left) and predicted drug targets (bottom right) and against Gene Ontology annotation **(C, D)** for biological process (GOBP), cellular component (GOCC) and molecular function (GOMF) categories. Radial distance on each radar plot corresponds to significance of enrichment. The most enriched feature in each comparison is indicated. Significance, radial axis, = $-\log_{10}$ adjusted P.

Fig. S6

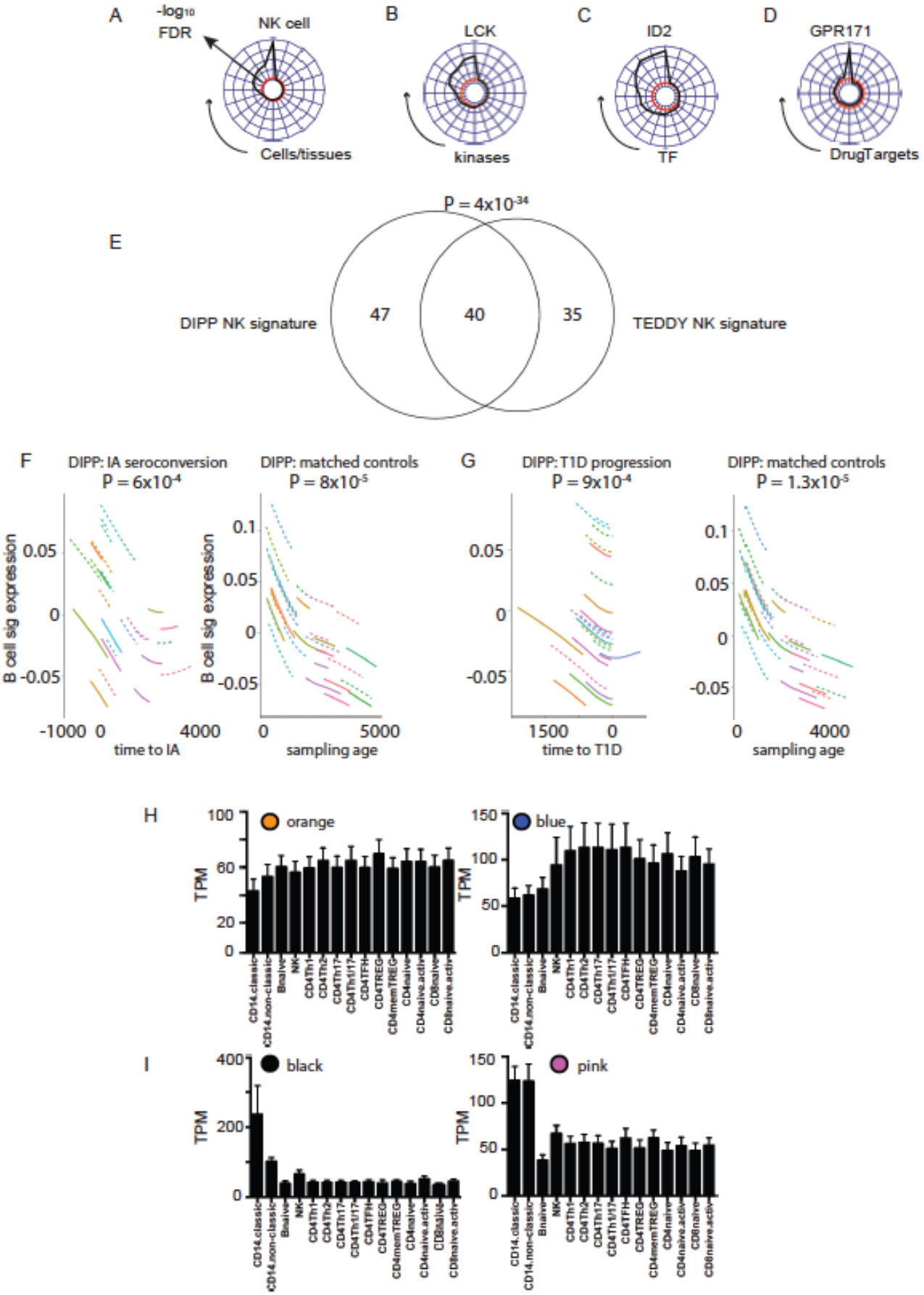
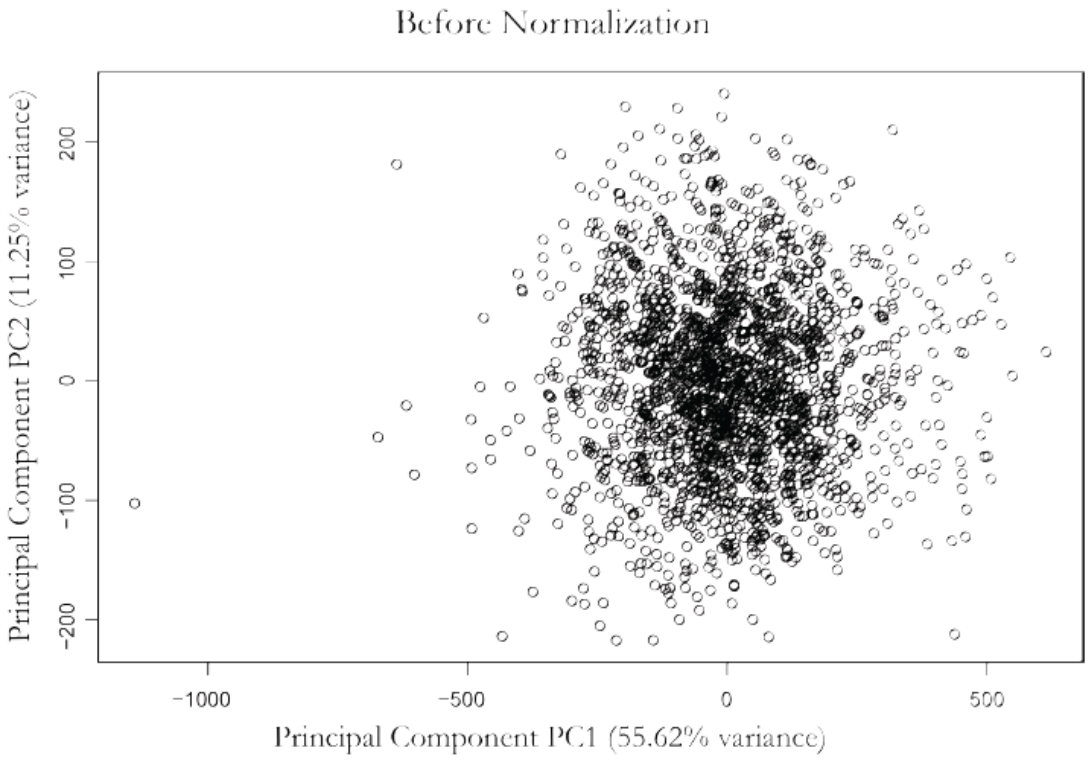
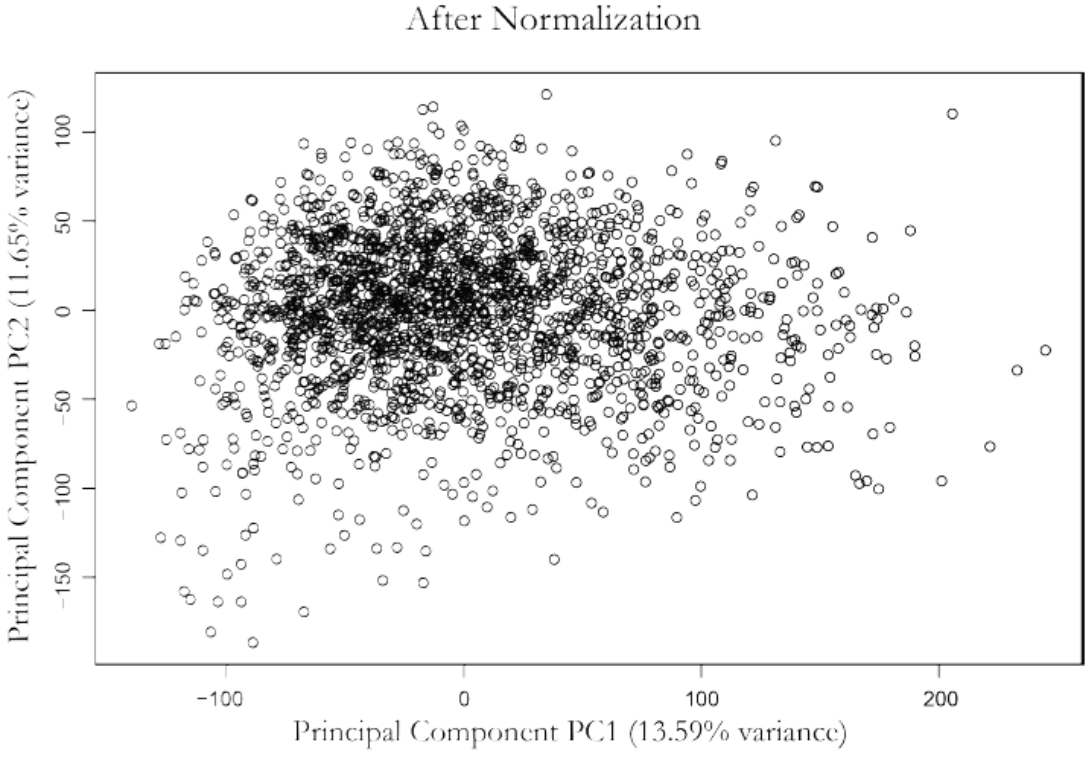

A
-log10 FDR
NK cell
Cells/tissues
B
LCK
kinases
C
ID2
TF
D
GPR171
DrugTargets
E
P = 4x10-34
DIPP NK signature
47
40
35
TEDDY NK signature
F
DIPP: IA seroconversion
P = 6x10-4
DIPP: matched controls
P = 8x10-5
B cell sig expression
time to IA
sampling age
G
DIPP: T1D progression
P = 9x10-4
DIPP: matched controls
P = 1.3x10-5
time to T1D
sampling age
H
orange
blue
TPM
I
black
pink
TPM

Figure S6. DIPP module enrichment. (A – D) Radar plots showing enrichment of the DIPP NK module (Fig. 3 J, K, n=87 probes) compared to the ARCHS4 repository of cell-type specific transcripts **(A)**, kinases **(B)**, transcription factors **(C)** and potential drug targets **(D)** as shown for the TEDDY NK signature in Fig. 2E. Radial distance from the centre = $-\log_{10}$adjustedP, red line = threshold FDR5%. **(E)** Venn diagram showing the overlap of genes included in the TEDDY NK (n=75) and DIPP NK (n=87) signatures, P = hypergeometric test. **(F-G)** Line plots showing individual effects of natural cubic spline-fitted linear mixed model for a B cell signature against time to IA seroconversion (**F**, left, n=183 samples from n=27 subjects) or age in controls (**F**, right, n=173 samples from n=31 subjects) and time to T1D onset (**G**, left ,n=122 samples from n=22 subjects) or age in controls (**G**, right, n=189 samples from n=34 subjects). **(H, I)** Barplots illustrating expression of each of the orange, blue, black and pink modular signatures against immune cell types represented in the DICE relational database. Y-axis, transcripts per million (TPM).

Fig.S7: Transcriptomic QC and batch correction





FigS7: Transcriptomic QC and batch correction.

A. First two principal components of the gene expression data before normalization.

B: First two principal components of the gene expression data after normalization.